Some sequential Monte Carlo techniques for Data Assimilation in a plant growth model

Yuting Chen, Samis Trevezas, Aman Gupta, Paul-Henry Cournède

SOME SEQUENTIAL MONTE CARLO TECHNIQUES FOR DATA ASSIMILATION IN A PLANT GROWTH MODEL

Yuting Chen[1], Samis Trevezas[1], Aman Gupta[1], and Paul-Henry Cournède[1]

Digiplante, Laboratory of Applied Mathematics and Systems, Ecole Centrale Paris, France
(E-mail: yuting.chen@ecp.fr)

Abstract. Data assimilation techniques have received considerable attention due to their capability to improve prediction and the most important applications concern weather forecasting and hydrology. Among many competing data assimilation approaches, those based on sequential Monte Carlo (SMC) methods, known as "particle filters", have gained their popularity because they are adaptive to nonlinearity and non-Gaussianity. In this study we test the performance of three SMC methods to predict biomass production and allocation in a dynamically evolving plant-growth model that can be formalized as a nonlinear state space model. The first method concerns a post-regularized particle filter (Musso and Oudjane[18]) which uses a mixture of Gaussian kernels (or more generally a kernel based method) to avoid sample impoverishment in the resampling step, the second and the third method involves the unscented Kalman filter (UKF) and the ensemble Kalman filter (EnKF) which are the extensions of classic Kalman filter invented for nonlinear systems. All the three approximate Bayesian estimation techniques deal simultaneously in their state vector fixed model parameters and state variables. We show that these methods perform well in realistic scenarios with sparse observations and discuss their limitations. Outside the context of data assimilation, we also present a maximum likelihood approach based on a stochastic version of an EM (Expectation-Maximization) algorithm, where the E-step can be approximated by the aforementioned SMC methods and discuss the pros and cons of the resulting algorithm. The performance of our methods is illustrated on data from the sugar-beet.

Keywords: plant growth model, data assimilation, sequential Monte-Carlo methods, kernel based method, stochastic EM algorithms, sugar-beet. .

1 Introduction

Due to inherent limitations in both the measurements and plant growth models, as measurements are often limited and unevenly distributed over time and the models are usually built based on assumptions coupled with some principles which inevitably lead to imperfectly defined parameters, the parametrization and the prediction of plant growth models have commonly been regarded as complex and critical issues. Sequential data assimilation techniques, especially the filtering methods have therefore received considerable attention, not only for the possibility to reconstruct the system by estimating simultaneously model parameters and state variables but also for their capability of identify

the sources of uncertainties in order to improve prediction accuracy and to reduce the corresponding confidence interval.

In a Baysian framework, the filtering methods provide distributional estimates instead of point estimates as most of the frequentist methods do, so while using a historical batch of data to fit the system, the filtering methods could take into account the variation of model parameters over time. This feature corresponds exactly the need of the prediction problem in the context of plant growth model, since it permits us to evaluate the uncertainty related to the estimated parameters of the model and to assess properly the uncertainties stemming from other sources in order to preserve them during the data assimilation step. However, the most desired the feature still remains to be the capacity of predicting nonlinear growth behaviours, especially to predict the occasionally occurred skewness due to the sudden or unusual climate changes.

When dealing with linear systems, the most efficient filtering method is Kalman filter (Kalman[13]). To date, many efforts have been made to develop extension for Kalman filter in nonlinear systems, the most well known extensions remain to be the extended Kalman filter, the unscented Kalman filter and the ensemble Kalman filter. The Extended Kalman Filter (EKF) (Evensen[9]) simply linearizes locally the model so that the tradition Kalman filter can be applied. However, in the case that the nonlinearity is significant, it may cause divergence and the method proves to be no longer reliable. On the other hand, the Unscented Kalman Filter (UKF) (Julier and Uhlmann[11], Quach *et al.*[21]) adopts deterministic sampling aiming at using a small set of discretely sampled points, known as sigma-points (Julier *et al.*[12], Wan and Van Der Merwe[22]), to get hold of the information of higher order for both mean and covariance matrix. The Ensemble Kalman Filter (EnKF) (Evensen[9]) relies on normality assumptions in order to improve the accuracy of its estimates with a more important number of samples compared to the UKF. Both latter methods generalize elegantly to nonlinear systems which are free of the linearization required by the EKF.

Another important alternative is Particle Filter (Gordon*et al.*[10], Kitagawa[14]), also known as Sequential Importance Sampling. Unlike the Kalman filter based methods, these Monte-Carlo filtering methods intend to provide better approximation of the exact posterior distributions by creating a set of randomly drawn samples with each an associated weight to present the probability distribution conditioned on a series of observations. However, their main weakness is the potential degeneracy (Arulampalam *et al.*[1]) and impoverishment (Gordon *et al.*[10]). With the purpose of alleviating the undesirable side effect of resampling to improve the parametrization performance when facing restricted dataset, we opted for the Convolution Particle Filter (CPF) proposed by Campillo and Rossi[3] based on the post-Regularized Particle Filter (post-RPF) (Musso and Oudjane[18], Oudjane and Musso[20], Le Gland and Oudjane[15]).

In this paper, we aim to investigate the properties and the performance of the three parameter estimation methods: the UKF, the EnKF and the CPF in the context of sequential data assimilation problems for a plant growth model with constraint observations, specifically regarding their abilities to provide

reliable a priori estimates for data assimilation and prediction purposes. The LNAS model of sugar beet growth (Cournède *et al.*[8]), as well as three years of experimental data obtained in comparable but different situations are employed in the application. One dataset is used for parameter estimation and the two others are used to test model prediction, with assimilation of the data from early growth stages.

In Section 2, the principles of the three filtering methods are recalled and their iterative version for parameter estimation is introduced. Section 3 contains a brief outline of the LNAS model and the experimental datasets, the description of the calibration and assimilation procedures are equally given. The results based on the real experimental data are displayed and the corresponding discussion is carried out in Section 4. Finally, some conclusions are presented in Section 5.

2 Methods

Plant growth models or crop models are generally deterministic and written in a state-space form, by introducing modeling and measurement noises, these models can also be elaborated to be stochastic. In both situations, they can cope with the statistical framework of sequential data assimilation designed to estimate the time evolving state variables, such as the yield. The difficulty for an appropriate data assimilation implementation lies not only on the uneven and irregular measurement data problems, but also on the robustness of predicting the occasionally occurred skewness due to the sudden or unusual climate changes.

The proposed approach consists of three steps. In the first place, the least influential model parameters are screened using sensitivity analysis methods (Campolongo *et al.*[4]) and are thus fixed. Regarding the fact that no satisfactory distributions are available for the selected parameters, which are believed to be the most influential ones, to perform the data assimilation directly, a first calibration is carried out subsequently based on a given experimental dataset in the second step. In this paper, the chosen model parameters are estimated with the three filtering methods with the same prior distribution. To obtain more precise estimations, the filtering process is iterated by taking the posterior distribution of iteration k as prior distribution for iteration $k+1$ until the convergence of all the estimates is achieved. The modeling and measurement noises can thereafter be evaluated. Meanwhile, since the final distributions are influenced by the regularization effect and by the empirical estimation of the modeling and measurement noises, they therefore can no longer represent the uncertainty of the estimates. The uncertainty related to the unknown parameters are hence assessed by parametric bootstrap.

During the third assimilation phase, the CPF approach is implemented again with the three sets of prior distributions provided by three filtering methods in the previous calibration step. A new comparable experimental dataset with few measurements is introduced, so that a recalibration can be carried out. The probability density is represented by a great number of samples (particles) which evolve in time. Model parameters and state variables are adjusted and

updated based on the available data of early growth stages. The predictions are then calculated based on the forecasted values of all the particles.

In this section, the general state-space model framework is presented. The three filtering algorithms as well as their iterative version are thereby briefly described in order to provide an outline of the parameter estimation step.

2.1 General State-Space Models:

Let a general nonlinear dynamic system be described by the following discrete time equations:

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

$X(t)$ represents the system state variable vector at time t, f operates the propagation of the model states. Θ is a vector of parameters of dimension p and $\eta(t)$ is the modeling noise, corresponding to model imperfections or uncertainty in the model inputs. $Y(t)$ is the noisy observation vector of the system which consists of state variables that can be observed experimentally and usually differ from $X(t)$ (such as biomasses of some plant organs that can be measured while the daily biomass production cannot). g is the transition operator which links the observations to the system states by adding measurement noises, 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: $X_n := X(t_n)$, $Y_n := Y\left(t_n\right)$ and $Y_{1:n} := \left(Y\left(t_1\right), Y\left(t_2\right), \ldots, Y\left(t_n\right)\right)$.

The objective of the filtering methods is to estimate jointly the parameters and the hidden states of the dynamic system by processing the data online. An augmented state vector $X_n^a = (X_n, \Theta_n)$ is thus defined with X_n the true hidden state at time t_n and Θ_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 $p(x_0^a)$.

2.2 Convolution Particle Filter

Particle filter has been regarded as a standard technique for performing recursive nonlinear estimation (Arulampalam *et al.*[1]). However, since the discrete approximation of the filtering distribution may result in sample impoverishment, a regularization strategy was invented to transform the discrete approximation to a continuous one, this approach is thus named as Post-Regularized Particle Filter (Oudjane and Musso[19]). 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 (Campillo and Rossi[3]), the approach is thus called the Convolution Particle Filter.

In the initialization step, the particles are initialized from either informative distributions ($p(x_0^a)$) or non-informative distributions. Uniform weights are assigned to each particle. Only at time steps when the observation is available that the filtering process is carried out with two steps:

Prediction: A kernel estimator denoted by $\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n})$ is built. 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 integrated forward in time by the evolution model until the next available measurement date to obtain the forecasted states $\{\tilde{x}_{n+1^-}^{a\,(i)}, i = 1, \ldots, M\}$. A weight is assigned to each particle based on the experimental measurements and the forecasted value. 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 $K_{h_M^X}^X$, with bandwidth parameter h_M^X:

$$\begin{aligned}\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n}) = \frac{1}{M}\sum_{i=1}^{M} K_{h_M^X}^X \left(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}\right) \\ \cdot\, p\left(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)}\right).\end{aligned} \tag{2}$$

Correction: The regularization is performed on the weighted samples, therefore the kernel approximation for $p(x_{n+1}^a|y_{1:n+1})$ can be expressed under the form:

$$\begin{aligned}\hat{p}(x_{n+1}^a|y_{1:n+1}) = \frac{1}{\sum_{i=1}^{M} p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})} \\ \cdot \sum_{i=1}^{M} K_{h_M^X}^X (x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}) p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)}).\end{aligned} \tag{3}$$

Where $p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})/\sum_{i=1}^{M} p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})$ can be regarded as the normalized weight $\tilde{w}_{n+1}^{(i)}$ associated to the particle $\tilde{x}_{n+1^-}^{a\,(i)}$. In the case that the likelihood function cannot be compute, inspired by the Post-Regularized Particle Filter, a convolution kernel is introduced to regularize the likelihood of the observation.

2.3 Unscented Kalman Filter

The Unscented Kalman Filter is known as one of the nonlinear extensions of classical Kalman Filter. When f_n^a and g_n are no longer linear, the density $p(x_n^a|y_{0:n})$ doesn't follow the normal distribution any more. Based on normal assumptions, an approximation is adopted by creating a series of sigma-points for nonlinear systems.

Prediction:

$d_\eta = dim(\eta_{n+1})$, $d_X = dim(\hat{x}_{n|n}^a)$ and $d_{\eta,X} = d_\eta + d_X$

J_n and R_n are the covariance matrix for η_n and ξ_n respectively

Compute the $2d_{\eta,X}+1$ sigma-points $\chi^i_{n|n}$ and their weights ω_i according to $\mathcal{N}\left(\hat{x}^b_{n|n}, \hat{\Sigma}^{x^b}_{n|n}\right)$, with

$$\hat{x}^b_{n|n} = (\hat{x}^a_{n|n}, \mathbf{0}_{d_\eta}) \qquad \text{and} \qquad \hat{\Sigma}^{x^b}_{n|n} = \begin{pmatrix} \hat{\Sigma}^{x^a}_{n|n} & \mathbf{0}_{d_X,d_\eta}. \\ \mathbf{0}_{d_\eta,d_X} & J_{n+1} \end{pmatrix}$$

We propagate the sigma-points to obtain the expectation at time $n+1$:

$$\hat{x}^a_{n+1|n} = \sum_{i=1}^{2d_{\eta,X}+1} \omega_i \chi^i_{n+1|n} \qquad \text{and} \qquad \hat{y}_{n+1|n} = \sum_{i=1}^{2d_{\eta,X}+1} \omega_i g_{n+1}(\chi^i_{n+1|n})$$

The associated covariance matrix :

$$\hat{\Sigma}^{x^a}_{n+1|n} = \sum_{i=1}^{2d_{\eta,X}+1} \omega_i (\chi^i_{n+1|n} - \hat{x}^a_{n+1|n})^T (\chi^i_{n+1|n} - \hat{x}^a_{n+1|n}). \tag{4}$$

$$\hat{\Sigma}^{y}_{n+1|n} = \sum_{i=1}^{2d_{\eta,X}+1} \omega_i (\zeta^i_{n+1|n} - \hat{y}_{n+1|n})^T (\zeta^i_{n+1|n} - \hat{y}_{n+1|n}). \tag{5}$$

$$\hat{\Sigma}^{x^a y}_{n+1|n} = \sum_{i=1}^{2d_{\eta,X}+1} \omega_i (\chi^i_{n+1|n} - \hat{x}^a_{n+1|n})^T (g_{n+1}(\chi^i_{n+1|n}) - \hat{y}_{n+1|n}). \tag{6}$$

Correction :
Compute the Kalman gain:

$$K_{n+1} = \hat{\Sigma}^{x^a y}_{n+1|n} \left(\hat{\Sigma}^{y}_{n+1|n}\right)^{-1}.$$

The corrected estimator and the corresponding covariance matrix at time $n+1$:

$$\hat{x}^a_{n+1|n+1} = \hat{x}^a_{n+1|n} + K_{n+1}(y_{n+1} - \hat{y}_{n+1|n})^T \tag{7}$$

$$\hat{\Sigma}^{x^a}_{n+1|n+1} = \hat{\Sigma}^{x^a}_{n+1|n} - K_{n+1}\hat{\Sigma}^{y}_{n+1|n}K^T_{n+1}. \tag{8}$$

2.4 Ensemble Kalman Filter

The Ensemble Kalman Filter is another extension of Kalman filter designed for nonlinear system. It's established based on the Monte-Carlo method coupled with the Kalman formulation.
Prediction:
The expectation of X^a_{n+1} given $Y_{0:n}$ can be obtained with the evolution equation:

$$\hat{x}^a_{n+1|n} = E[X^a_{n+1}|Y_{0:n}] = E[f^a_{n+1}(X^a_n, \eta^a_{n+1})|Y_{0:n}].$$

The covariance matrix associated to X^a_{n+1} given $Y_{0:n}$ is:

$$\hat{\Sigma}^{x^a}_{n+1|n} = E[(X^a_{n+1} - \hat{x}^a_{n+1|n})^T (X^a_{n+1} - \hat{x}^a_{n+1|n})|Y_{0:n}]. \tag{9}$$

In the same way for $p(y_{n+1}|y_{0:n})$, we may obtain the corresponding expectation as following:

$$\hat{y}_{n+1|n} = E[Y_{n+1}|Y_{0:n}] = E[g_{n+1}(X^a_{n+1}) + \xi_{n+1}|Y_{0:n}] = E[g_{n+1}(X^a_{n+1})|Y_{0:n}] \tag{10}$$

and the associate covariance matrix can thus be calculated:

$$\begin{aligned}\hat{\Sigma}^y_{n+1|n} &= E[(Y_{n+1} - \hat{y}_{n+1|n})^T(Y_{n+1} - \hat{y}_{n+1|n})|Y_{0:n}] \\ &= E[(g_{n+1}(X^a_{n+1}) - \hat{y}_{n+1|n})^T(g_{n+1}(X^a_{n+1}) - \hat{y}_{n+1|n})|Y_{0:n}] + R_{n+1}.\end{aligned} \tag{11}$$

The cross correlation matrix of X^a_{n+1} and Y_{n+1} given $Y_{0:n}$ is also corrected as follows:

$$\hat{\Sigma}^{x^a y}_{n+1|n} = E[(X^a_{n+1} - \hat{x}^a_{n+1|n})^T(Y_{n+1} - \hat{y}_{n+1|n})|Y_{0:n}]. \tag{12}$$

Correction :
Computed in the same way as in the UKF approach.

2.5 Iterative Filtering and the Conditional Approach

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 filtering (Chen *et al.*[6]) can hence be applied. At iteration k, the particles $x^{a(i)}_0$ 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 $\{\tilde{\Theta}^{(i)}_0, i = 1, \ldots, M\}$ 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 $k-1$. An averaging technique (Cappé *et al.*[5]) is used to smooth parameter estimates after a small burn-in period.

In order to evaluate the modeling and the observation noises, a conditional maximization approach (Chen *et al.*[7]) is carried out to estimate the noise related parameters, denoted Θ_2. Based on the results of the model parameter (denoted Θ_1) and state variable estimation performed by the iterative filtering processes, the noise parameters are therefore estimated empirically. The noise parameter estimation and the model parameter estimation are alternated so as to provide coherence estimates of Θ.

3 Application

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 (Monteith[17]): $\left(1 - e^{-\lambda \cdot Q_g(t)}\right)$ represents the fraction of intercepted radiation, with λ ($g^{-1}.m^2$) 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 μ (in $g \cdot MJ^{-1}\cdot$):

$$Q(t) = \left(\mu \cdot PAR(t) \left(1 - e^{-\lambda Q_g(t)}\right)\right) \cdot (1 + \eta_Q(t)) \tag{13}$$

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

Allocation for the foliage and root system compartments:

$$Q_f(t+1) = \quad Q_f(t) + \gamma(t) \cdot Q(t) \tag{14}$$

$$Q_r(t+1) = Q_r(t) + (1 - \gamma(t)) \cdot Q(t) \tag{15}$$

where

$$\gamma(t) = (\gamma_0 + (\gamma_f - \gamma_0) \cdot G_a(\tau(t))) \cdot (1 + \eta_\gamma(t)) \tag{16}$$

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 μ_a and standard deviation s_a, and the modeling 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 μ_s and standard deviation s_s:

$$Q_s(t) = G_s(\tau(t) - \tau_{sen}) Q_f(t) \tag{17}$$

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

$$Q_g(t) = Q_f(t) - Q_s(t) \tag{18}$$

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

$$Y(t) = \begin{pmatrix} Q_g(t) \cdot (1 + \epsilon_g(t)) \\ Q_r(t) \cdot (1 + \epsilon_r(t)) \end{pmatrix} \tag{19}$$

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 concerned datasets consist of limited experimental observations furnished by the French institute for Sugar Beet research (ITB, Paris, France) in 2006, 2008 and 2010 with different cultivars and measured in different locations (further experimental protocols details are presented in Lemaire *et al.*[16]). The 2010 dataset is chosen for calibration simply since it contains more observation points. Dry matter of root and leaves of 50 plants were collected at 14 dates:

$$\mathcal{O}_{2010} = \{54, 68, 76, 83, 90, 98, 104, 110, 118, 125, 132, 139, 145, 160\},$$

whilst for the assimilation and prediction step, the two other datasets (2006 and 2008) are used for which the same type of observations, the green foliage mass denoted by Q_g and the root compartment mass denoted by Q_r, were made but only at 7 dates:

$$\mathcal{O}_{2006} = \{54, 59, 66, 88, 114, 142, 198\},$$

$$\mathcal{O}_{2008} = \{39, 60, 67, 75, 88, 122, 158\}.$$

The final observations contains the mean value calculated based on all the samples and extrapolated at m^2.

3.3 Three-step analysis for prediction

The prediction process is carried out in three steps as indicated in Fig. 1.

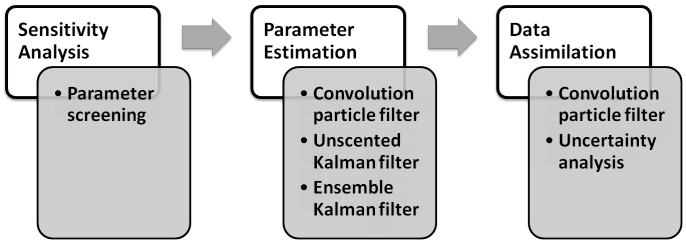

Fig. 1. The flow of the three-step analysis for prediction.

Parameter Screening by Sensitivity Analysis: Plant growth models contains frequently a large number of parameters which make the parametric estimation from experimental data difficult with important uncertainty linked to the estimates. With the purpose of selecting parameters identified as the most influential ones to be estimated, sensitivity analysis is therefore applied. Those screened as the least influential parameters can be fixed to any values in their domains. This method is called "screening" or "factor fixing" (Campolongo *et al.*[4]).

With this objective, we use the algorithm proposed by Wu *et al.*[23] 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. 2, we screen the parameters s_a, μ_{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 suggest that they should be estimated from experimental data.

Parameter Estimation: Based on the sensitivity analysis results, the unknown parameter vector for the deterministic part of the model is $\Theta_1 = (\mu, \lambda, \mu_a, \gamma_0, \gamma_f)$

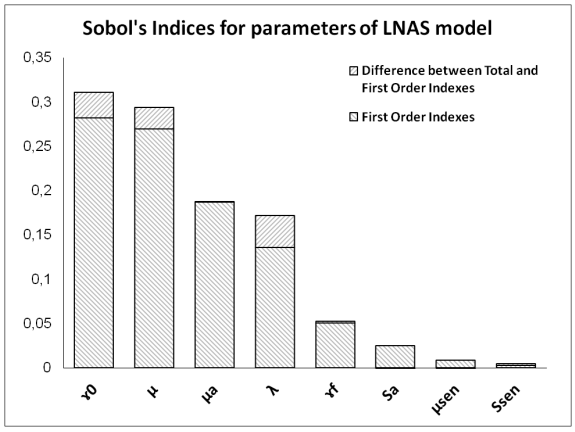

Fig. 2. Comparison of the first and total order indexes for μ, λ, γ_0, γ_f, μ_a, s_a, μ_{sen} and s_{sen}.

and the unknown noise parameter vector is $\Theta_2 = (\sigma_Q, \sigma_\gamma, \sigma_g, \sigma_r)$. For the conditional ICPF approach, 150000 particles were initialized with the same prior distributions as for the conditional IEnKF and IUKF approach. For the conditional IEnKF approach, an ensemble size of 400 is adopted. For all the three approaches, the conditional estimation process began with the estimation of Θ_1 given Θ_2, then Θ_2 was estimated empirically based on the estimates of the hidden states. The estimation then proceeded with the new value of Θ_2 and iterated. After the convergence of both Θ_1 and Θ_2, which is claimed by a standard stopping rule based on the relative changes in the estimations from three successive estimations (Booth and Hobert[2]). A parametric bootstrap was also achieved to evaluate the estimates' uncertainty. Standard deviations and confidence intervals were hence obtained from 100 bootstrap samples. The corresponding results are given in Table 1.

Data Assimilation with CPF: In the previous calibration step, all of the three filtering methods are applied to the LNAS model which allow us to estimate jointly the unknown parameters and the hidden state variables base on a historical batch of data (2010 dataset). The parametric bootstrap results provided by the three filtering methods were used as prior information of the CPF method in the assimilation step. 500000 particles were simulated for prediction purpose. For both the 2006 and 2008 datasets, all but the last two measurements were used to update the parameter and state estimates. Regarding the 2006 dataset, after day 114 (resp. day 88 for the 2008 dataset), the propagation of particles continued without any further correction. 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 (second phase of calibration based on the data of early growth stages), Uncertainty Analysis (UA) is also performed. 500000 simulations were initialized in

Parameter	IEnKF		ICPF		IUKF	
	Estimates	Std.	Estimates	Std.	Estimates	Std.
μ	3.60	0.15	3.56	0.12	3.90	0.27
λ	60.16	6.51	59.55	3.13	60.18	6.49
γ_0	0.83	0.09	0.84	0.04	0.80	0.06
γ_f	0.206	0.058	0.194	0.053	0.216	0.048
μ_a	639.39	83.28	642.33	62.40	579.98	73.75
s_a	276.18	149.83	308.69	109.95	338.31	44.89
σ_Q	0.040	-	0.042	-	$0.040^\star$	-
σ_γ	0.061	-	0.064	-	$0.060^\star$	-
σ_g	0.137	-	0.142	-	0.156	-
σ_r	0.166	-	0.165	-	0.193	-
Likelihood	-167.375	-	-164.733	-	-176.685	-
AIC	354.750	-	349.466	-	369.370	-
BIC	368.072	-	362.788	-	380.028	-

Table 1. Estimated values and approximated standard deviations for the IEnKF, ICPF and IUKF estimation for 6 functional parameters and 4 noise parameters of LNAS model. *:For the IUKF method, due to the estimation limitation, the modeling noise parameters σ_Q and σ_γ are fixed based on the estimation given by the two other methods.

the same way as in the CPF approach, which indicates that samples of Θ_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

Better performance of ICPF estimates are noted when it comes to the evaluation of log-likelihood as well as the AIC and AICc criteria during the calibration step based on the 2010 dataset, as illustrated by Table 1.

Generally speaking the assimilation step has well established its undeniable value as demonstrated by Table 2 compared to the prediction results without assimilation. The estimation relative error was reduced up to 42.4% when data assimilation was performed. In the meantime, the standard error related to the prediction was also significantly decreased in all cases when the second calibration has taken place based on the early growth data.

For green leaf biomass allocation, it has always been the ICPF estimates that gave the best predictions for both years, as indicated by Fig. 6 and Fig. 7. This may be related to its important nonlinearity.

IUKF provided the best 2008 root prediction, for the root biomass allocation is more linear then the green leaf biomass allocation according to Fig. 3 and Fig 4. This may also explain the fact that the prediction based on the IUKF estimates for the green leaf biomass are less accurate compared to the results given by the other two estimates.

However, it is not quite the case for the 2006 root biomass prediction. Although the IEnKF estimates showed best performance, we notice that the

		IEnKF		ICPF		IUKF	
		DA	UA	DA	UA	DA	UA
$Q_b\,(t_{142})$	Relative error	4.2%	45.8%	2.1%	44.5%	6.1%	55.6%
	Std.	56.0	163.7	56.1	128.7	62.7	165.0
$Q_b\,(t_{198})$	Relative error	11.2%	46.5%	7.2%	43.4%	14.4%	58.5%
	Std.	63.1	141.2	61.5	116.8	68.3	144.2
$Q_r\,(t_{142})$	Relative error	4.8%	29.8%	5.7%	32.2%	6.3%	43.1%
	Std.	256.0	384.3	256.1	354.3	303.4	479.2
$Q_r\,(t_{198})$	Relative error	2.5%	2.5%	3.9%	20.6%	4.8%	29.8%
	Std.	420.4	553.4	419.9	522.3	503.7	693.6

Table 2. Comparison of model prediction capacity of estimates provided by three filtering methods (IEnKF, ICPF and IUKF) based on the 2006 dataset with and without assimilation of data at the early growth stage. DA: with data assimilation, UA: uncertainty analysis without data assimilation.

performance of the three sets of estimates were relatively close, as suggested by Fig. 5.

Therefore in general, the most accurate predictions are still provided by the ICPF, regardless the time and the memory it required.

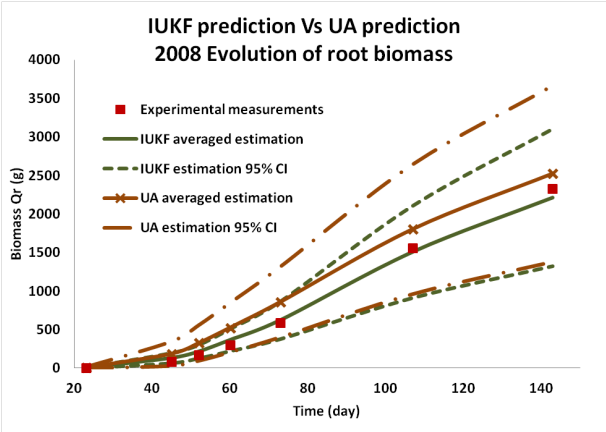

Fig. 3. Comparison of the predictions of Q_r in 2008 performed with or without data assimilation (Uncertainty Analysis) based on IUKF estimates.

Considering the time-consuming problem of the iterative version of particle filtering methods, it might be interesting to be less exigent on the model parameter estimation and loosen the convergence criterion so as to reduce the iteration numbers. Given the fact that the parameters are to be adjusted in the assimilation step, the importance of the point estimation in the calibration step might be over evaluated. However, it is noteworthy that the noise related parameters have played an crucial role in the uncertainty assessment of the

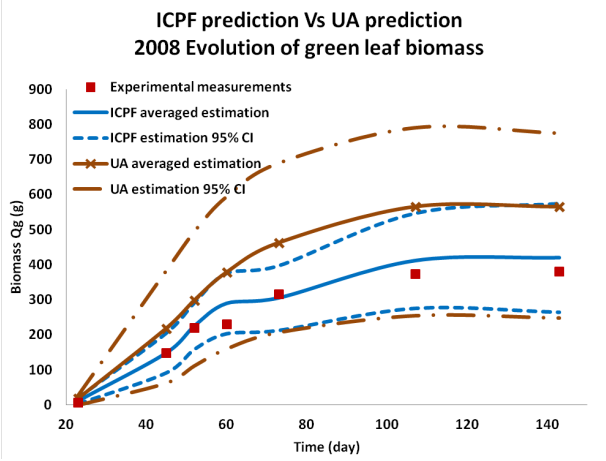

Fig. 4. Comparison of the predictions of Q_g in 2008 performed with or without data assimilation (Uncertainty Analysis) based on ICPF estimates.

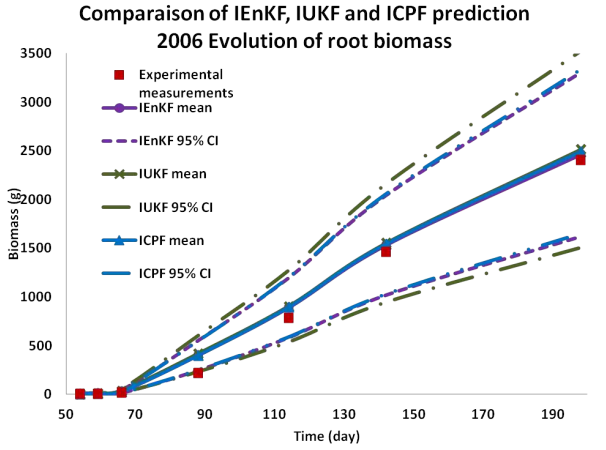

Fig. 5. Comparison of the predictions of Q_r in 2006 performed with data assimilation based on IEnKF, IUKF and ICPF estimates.

	Real Data 2006	IEnKF estimates (relative error in %)	Std.	ICPF estimates (relative error in %)	Std.	IUKF estimates (relative error in %)	Std.
$Q_b\,(t_{142})$	355.2	370.2 (4.2%)	56.0	362.7 (2.1%)	56.1	376.8 (6.1%)	62.7
$Q_b\,(t_{198})$	320.6	356.6 (11.2%)	63.1	343.8 (7.2%)	61.5	366.8 (14.4%)	68.3
$Q_r\,(t_{142})$	1459.2	1529.1 (4.8%)	256.0	1542.7 (5.7%)	256.1	1551.3 (6.3%)	303.4
$Q_r\,(t_{198})$	2400.0	2460.6 (2.5%)	420.4	2493.0 (3.9%)	419.9	2513.9 (4.8%)	503.7

Table 3. Comparison of model prediction capacity of estimates provided by three filtering methods (IEnKF, ICPF and IUKF) based on the 2006 dataset with assimilation of data at the early growth stage.

predictions. According to our tests, the prediction results are very sensitive to the level of observation noises. Since the level of noise parameter evaluated on one year cannot represent the level of observations made in another year

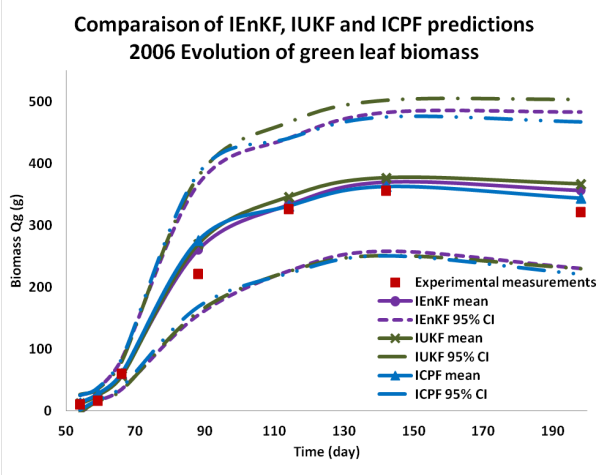

Fig. 6. Comparison of the predictions of Q_g in 2006 performed with data assimilation based on IEnKF, IUKF and ICPF estimates.

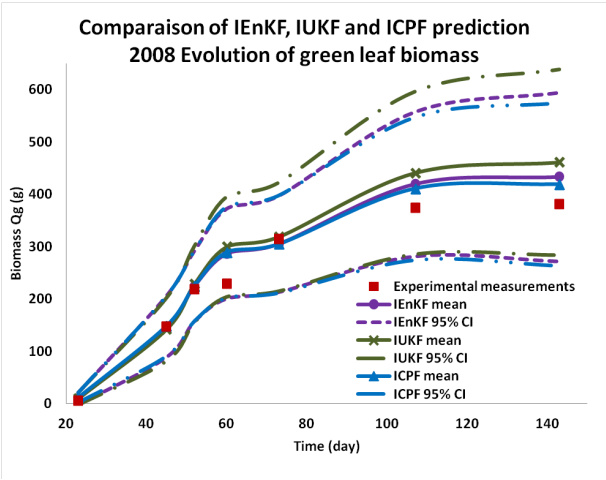

Fig. 7. Comparison of the predictions of Q_g in 2008 performed with data assimilation based on IEnKF, IUKF and ICPF estimates.

	Real Data 2008	IEnKF estimates (relative error in %)	Std.	ICPF estimates (relative error in %)	Std.	IUKF estimates (relative error in %)	Std.
$Q_b\,(t_{122})$	373.5	419.5 (12.3%)	68.9	410.6 (9.9%)	68.1	440.9 (18.1%)	77.4
$Q_b\,(t_{158})$	380.6	433.1 (13.8%)	80.5	418.9 (10.1%)	77.8	461.1 (21.2%)	88.4
$Q_r\,(t_{122})$	1559.1	1460.5 (6.3%)	248.5	1466.7 (5.9%)	246.7	1508.8 (3.2%)	298.3
$Q_r\,(t_{158})$	2327.7	2125.6 (8.7%)	367.1	2137.6 (8.2%)	363.2	2213.7 (4.9%)	444.4

Table 4. Comparison of model prediction capacity of estimates provided by three filtering methods (IEnKF, ICPF and IUKF) based on the 2006 dataset with assimilation of data at the early growth stage.

or in a different location, applying the same level of noises in the assimilation

step with a different dataset is debatable. Further studies are hence needed to address the influence of the observation noise level to the prediction quality.

5 Conclusion

Overall, the proposed three-step sequential data assimilation approach allows us to address properly various sources of uncertainties and to obtain satisfactory prediction results. The filtering methods allow us to consider that certain model parameters are time-variant. The procedure used in this study preserve the flexibility to investigate some possible time-variant parameters.

Regarding the three filtering methods, despite the normal assumption dependency, the IEnKF has provided a proper approximation for the nonlinear growth model LNAS. Although globally the ICPF provided the most accurate prediction and the smallest standard deviation, considering the fact that it is far more time-consuming compared to the IEnKF approach during the calibration step, the results suggest more advantages in the sequential data assimilation for the latter approach. Therefore, when the nonlinearity is not extremely important, the IEnKF is recommended. However, if the nonlinearity of the model is remarkable or in the case that cannot be easily evaluated, the ICPF approach is more suitable.

References

1.M.S. Arulampalam, S. Maskell, and T. Gordon, N.and Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. In *Connection between forest resources and wood quality: modelling approaches and simulation software*, number 50(2), page 1740188. IEEE Trans. Signal Proces., 2002.

2.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(1):265–285, 1999.

3.F. Campillo and V. Rossi. Convolution Particle Filter for Parameter Estimation in General State-Space Models. *IEEE Transactions in Aerospace and Electronics.*, 45(3):1063–1072, 2009.

4.F. Campolongo, J. Cariboni, and A. Saltelli. An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software*, 22:1509–1518, 2007.

5.O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models.* Springer, New York, 2005.

6.Y.T. Chen and P.-H. Cournède. Assessment of parameter uncertainty in plant growth model identification. In M. Kang, Y. Dumont, and Y. Guo, editors, *Plant growth Modeling, simulation, visualization and their Applications (PMA12).* IEEE Computer Society (Los Alamitos, California), 2012.

7.Y.T. Chen, S. Trevezas, and P.-H. Cournède. Iterative convolution particle filtering for nonlinear parameter estimation and data assimilation with application to crop yield prediction. In *Society for Industrial and Applied Mathematics (SIAM): Control & its Applications,San Diego, USA.*, 2013.

8.P.-H. Cournède, Y.T. Chen, Q.L. Wu, C. Baey, and B. Bayol. Development and evaluation of plant growth models: Methodology and implementation in the pygmalion platform. *Submitted*, 2013.

9.G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.*, (99(C5)):10143–10162, 1994.
10.N. Gordon, D. Salmond, and A.F.M Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Proc. Inst. Electr. Eng.*, Part F 140:107–113, 1993.
11.S. Julier and J. Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. *International Symposium of Aerospace/Defense Sensing, Simulation and Controls*, 1997. Orlando. Fl.
12.S. Julier, J. Uhlmann, and H. Durrant-Whyte. A New Method for the Non-Linear Transformation of Means and Covariances in Filters and Estimators. *IEEE Transaction on Automatic Control*, 45(3):477–482, 2000.
13.R.E. Kalman. A New Approach to Linear Filtering and Prediction Problem. *Journal of the basic engineering*, 82:35–45, 1960.
14.G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.
15.F. Le Gland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann. Appl. Probab.*, 14(1):144–187, 2004.
16.S. Lemaire, F. Maupas, P.-H. Cournède, J.-M. Allirand, P. de Reffye, and B. Ney. Analysis of the density effects on the source-sink dynamics in sugar-beet growth. In B.-G. Li, M. Jaeger, and Y. Guo, editors, *3rd international symposium on Plant Growth and Applications(PMA09), Beijing, China*. IEEE Computer Society (Los Alamitos, California), November 9-12 2009.
17.J.L. Monteith. Climate and the efficiency of crop production in britain. *Proceedings of the Royal Society of London B*, 281:277–294, 1977.
18.C. Musso and N. Oudjane. Regularization schemes for branching particle systems as a numerical solving method of the nonlinear filtering problem. In *Proceedings of the Irish Signals and Systems Conference*, Dublin, 1998.
19.N. Oudjane and C. Musso. Regularized particle schemes applied to the tracking problem. In *International Radar Symposium, Munich, Proceedings*, 1998.
20.N. Oudjane and C. Musso. Multiple model particle filter. In *17ème Colloque GRETSI, Vannes 1999*, pages 681–684, 1999.
21.M Quach, N. Brunel, and F. d'Alché Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.
22.E. Wan and R. Van Der Merwe. The Unscented Kalman Filter for Non-Linear Estimation. *IEEE Symposium 2000, lake Louise, Alberta, Canada*, 2000.
23.Q.L. 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*, 107:35–43, 2012.