CentraleSupélec

Reduction of parameter uncertainty and genotype differentiation in plant growth models

Paul-Henry Cournède, *et al.*

Digiplante, Lab of Mathematics and Computer Science, CentraleSupélec

HortiModel, Avignon, 21-09-2016

A Plant Science Issue : Interaction Genotype×Environment

Biophysical models can help understand and predict this interaction :
« 1 genotype = 1 stable parameter vector », [Tardieu, 2003]

A Plant Science Issue : Interaction Genotype×Environment

Biophysical models can help understand and predict this interaction :
« 1 genotype = 1 stable parameter vector », [Tardieu, 2003]

Objectives

- Phenotype = f1(Parameters, Environment)
- Parameter = f2(Genetics)

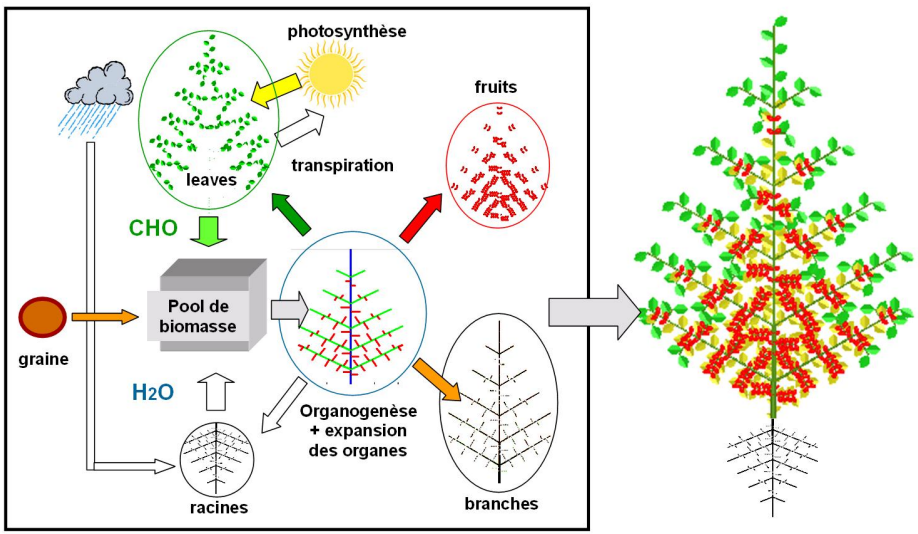
A Plant Science Issue : Interaction Genotype×Environment

Biophysical models can help understand and predict this interaction :
« 1 genotype = 1 stable parameter vector », [Tardieu, 2003]

Objectives

- Phenotype = f1(Parameters, Environment)
- Parameter = f2(Genetics)

Dynamic System of Plant Growth



$$X(t+1) = F(X(t), U(t), \theta, t)$$

- $X(t)$: state variables $\Longrightarrow$ organ masses, leaf surfaces...
- $F \Longrightarrow$ biophysical laws
- θ : parameters $\Longrightarrow$ genotype specific
- $U(t)$: exogeneous variables $\Longrightarrow$ environmental and cultural conditions

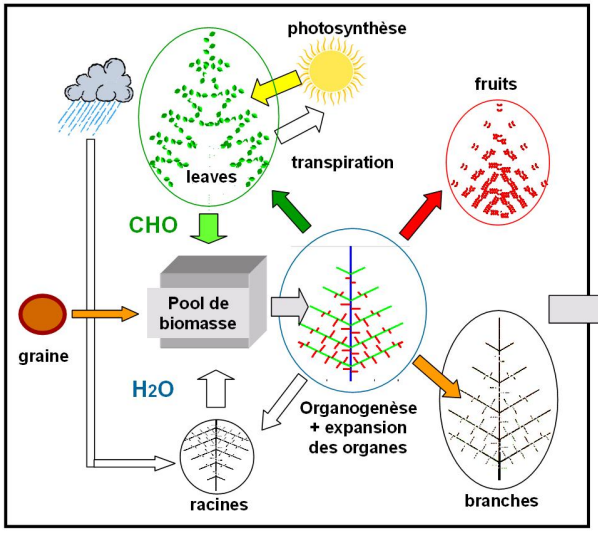
A Plant Science Issue : Interaction Genotype×Environment

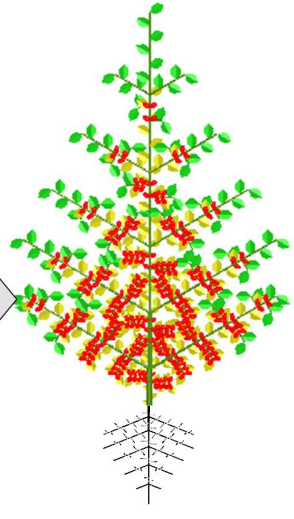
Biophysical models can help understand and predict this interaction :
« 1 genotype = 1 stable parameter vector », [Tardieu, 2003]

Objectives

- Phenotype = f1(Parameters, Environment)
- Parameter = f2(Genetics)

Dynamic System of Plant Growth





$$X(t+1) = F(X(t), U(t), \theta, t)$$

- $X(t)$: state variables $\Longrightarrow$ organ masses, leaf surfaces...
- $F \Longrightarrow$ biophysical laws
- θ : parameters $\Longrightarrow$ genotype specific
- $U(t)$: exogeneous variables $\Longrightarrow$ environmental and cultural conditions

$\Longrightarrow$ A heavy tendency : the development of more and more mechanistic models, with more and more processes (even multiscale processes), and more and more parameters.

A Methodological Issue : Model Parameterization

Different Methods

- Direct measurements
- Literature data
- Similar or comparable experiments

A Methodological Issue : Model Parameterization

Different Methods

- Direct measurements
- Literature data
- Similar or comparable experiments
- Hidden parameter estimation from experimental data (model inversion)

$\Longrightarrow$ Our preference is in all cases to run a full parameter estimation from experimental data, in order to assess properly parameter uncertainty... but it necessitates a proper statistical framework.

Parameter Estimation and Uncertainty Evaluation

Formulation of Plant State-Space Models as Hidden Markov Models

$$\begin{cases} X_{t+1}|X_t \sim p\left(x_{t+1}|x_t, \theta\right) \\ Y_t|X_t \sim p\left(y_t|x_t, \theta\right) \end{cases}$$

X_t : hidden variables, Y_t : observed variables, θ : unknown parameters.

Parameter Estimation and Uncertainty Evaluation

Formulation of Plant State-Space Models as Hidden Markov Models

$$\begin{cases} X_{t+1}|X_t \sim p\left(x_{t+1}|x_t, \theta\right) \\ Y_t|X_t \sim p\left(y_t|x_t, \theta\right) \end{cases}$$

X_t : hidden variables, Y_t : observed variables, θ : unknown parameters.

Maximum likelihood estimation

$$\hat{\theta} = Argmax\left(\mathcal{L}(\theta; y)\right) \ ,$$

with $\mathcal{L}(\theta; y) = p(y|\theta)$ via stoch. variants of the EM algorithm [Trevezas and C., 2013]

Parameter Estimation and Uncertainty Evaluation

Formulation of Plant State-Space Models as Hidden Markov Models

$$\begin{cases} X_{t+1}|X_t \sim p\left(x_{t+1}|x_t, \theta\right) \\ Y_t|X_t \sim p\left(y_t|x_t, \theta\right) \end{cases}$$

X_t : hidden variables, Y_t : observed variables, θ : unknown parameters.

Maximum likelihood estimation

$$\hat{\theta} = Argmax\left(\mathcal{L}(\theta; y)\right) \ ,$$

with $\mathcal{L}(\theta; y) = p(y|\theta)$ via stoch. variants of the EM algorithm [Trevezas and C., 2013]

Bayesian estimation

Evaluation of the posterior $p(\theta|y)$ from the prior $p(\theta)$, via MCMC or filtering methods)

Parameter Estimation and Uncertainty Evaluation

Formulation of Plant State-Space Models as Hidden Markov Models

$$\begin{cases} X_{t+1}|X_t \sim p\left(x_{t+1}|x_t, \theta\right) \\ Y_t|X_t \sim p\left(y_t|x_t, \theta\right) \end{cases}$$

X_t : hidden variables, Y_t : observed variables, θ : unknown parameters.

Maximum likelihood estimation

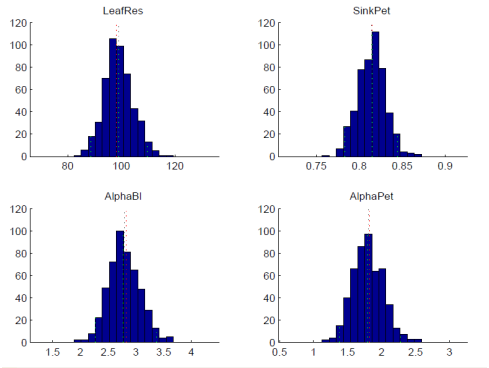
$$\hat{\theta} = Argmax\left(\mathcal{L}(\theta; y)\right) \ ,$$

with $\mathcal{L}(\theta; y) = p(y|\theta)$ via stoch. variants of the EM algorithm [Trevezas and C., 2013]

Bayesian estimation

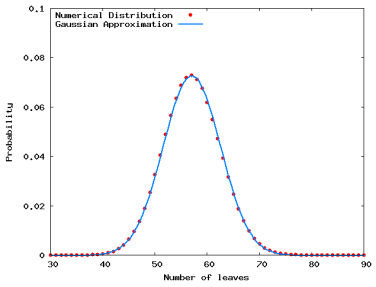
Evaluation of the posterior $p(\theta|y)$ from the prior $p(\theta)$, via MCMC or filtering methods)

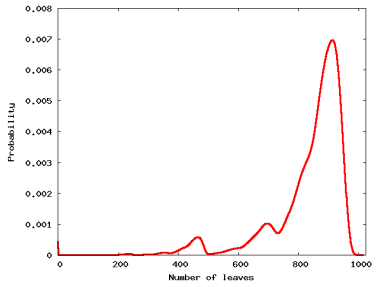
Resulting Distributions

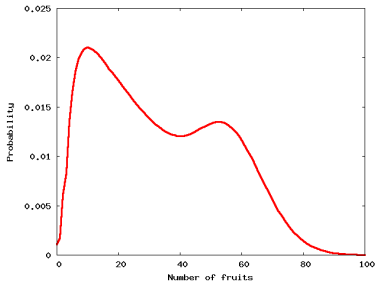


Difficulties Linked to Parameter Uncertainty

In terms of Prediction : Ucertainty Propagation

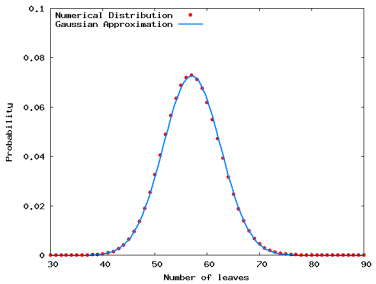
F (

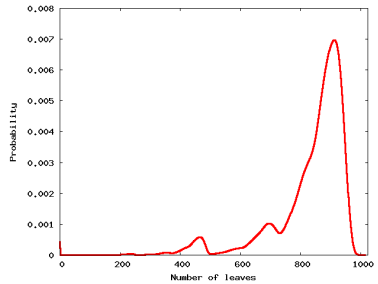
, 

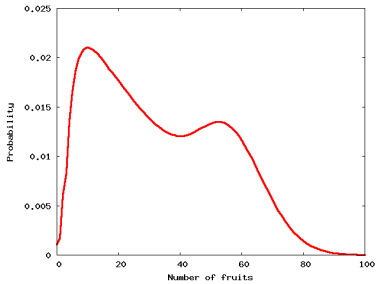
) = 

Difficulties Linked to Parameter Uncertainty

In terms of Prediction : Ucertainty Propagation

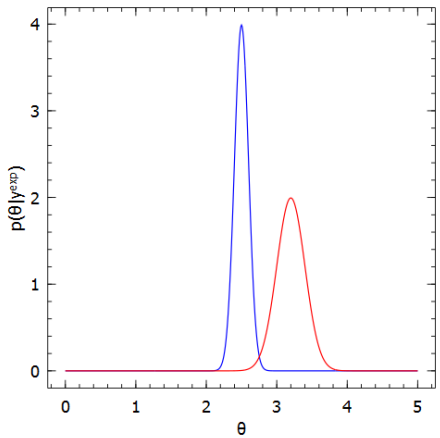
F (

, 

) = 

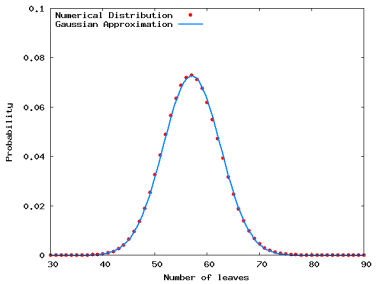
In terms of Genotype Differentiation

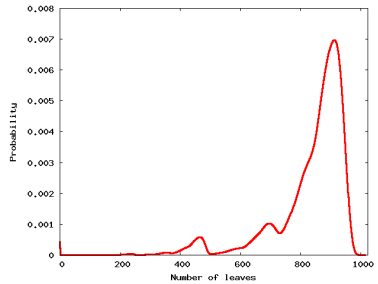
2 genotypes, A and B, $\hat{\theta}_A = 2.5$, $\hat{\theta}_B = 3.3$, with $p(\theta_A|y_A)$, $p(\theta_B|y_B)$

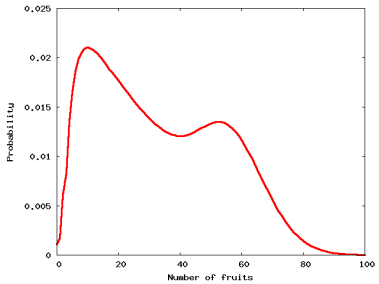


Difficulties Linked to Parameter Uncertainty

In terms of Prediction : Ucertainty Propagation

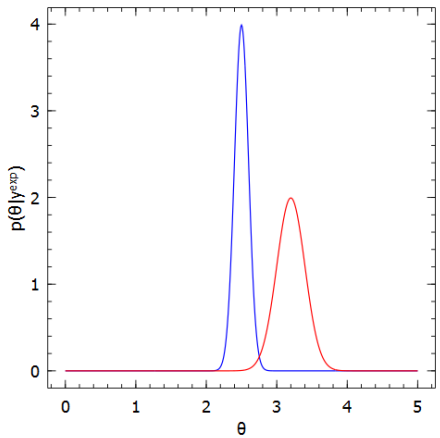
F (

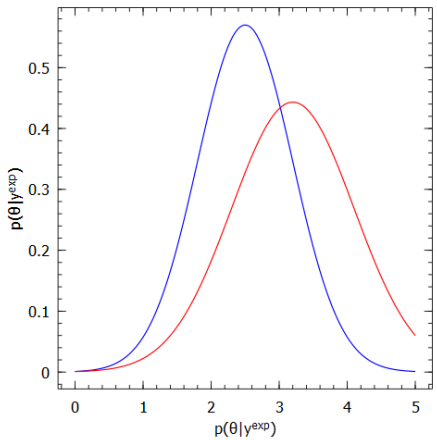
, 

) = 

In terms of Genotype Differentiation

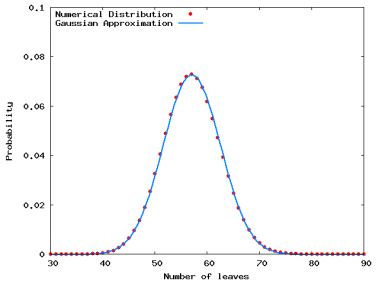
2 genotypes, A and B, $\hat{\theta}_A = 2.5$, $\hat{\theta}_B = 3.3$, with $p(\theta_A|y_A)$, $p(\theta_B|y_B)$

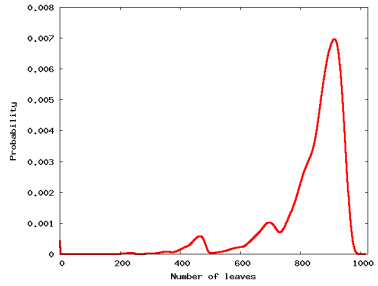


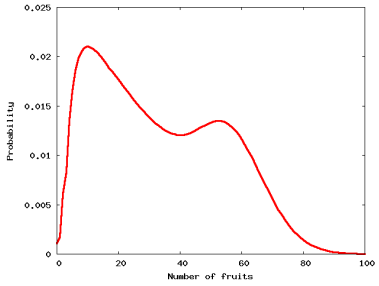


Difficulties Linked to Parameter Uncertainty

In terms of Prediction : Ucertainty Propagation

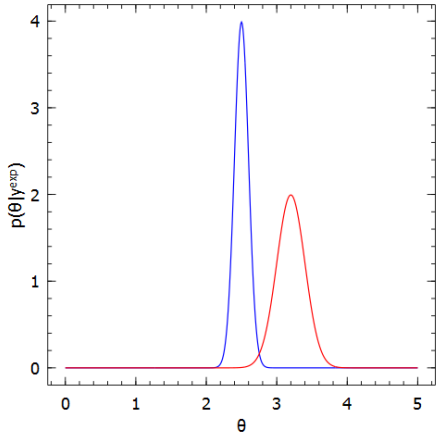
F (

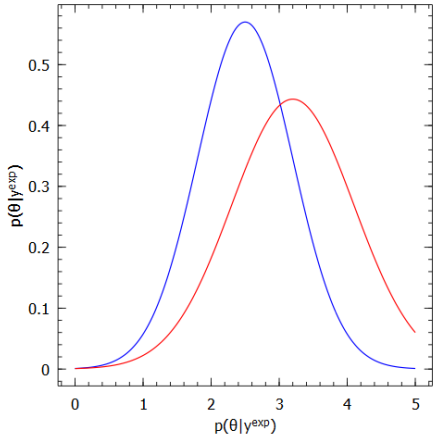
, 

) = 

In terms of Genotype Differentiation

2 genotypes, A and B, $\hat{\theta}_A = 2.5$, $\hat{\theta}_B = 3.3$, with $p(\theta_A|y_A)$, $p(\theta_B|y_B)$





$\Longrightarrow$ A problem of adequacy between model complexity and experimental data ?

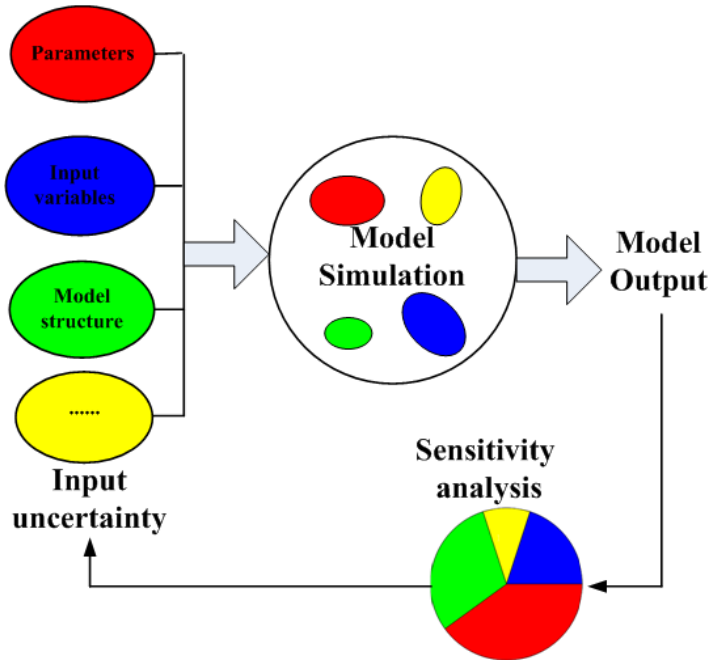
Outline

Outline

Global Sensitivity Analysis for Plant Growth Modeling

Sensitivity Analysis

'The study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs' [Saltelli et al.[2004]]



- Input factors $[X_i \ (1 \leq i \leq k)] \Longrightarrow$ described by random distributions
  - Uncertain parameters
  - Input variables
- Model execution $[f(\mathbf{X}_n) \ (1 \leq n \leq N)]$
- Output of interest $[\mathbf{Y} = f(\mathbf{X})] \Longrightarrow$ depends on analysis aims

Interest of Sensitivity Analysis in the Modeling Process

- To help for the **parameterization** of Plant Models :
  - Factor Priorization (FP) : identification of the most important factors
  - Factor Fixing (FF) : identification of the most non-influential factors (screening)

Interest of Sensitivity Analysis in the Modeling Process

- To help for the **parameterization** of Plant Models :
  - Factor Priorization (FP) : identification of the most important factors
  - Factor Fixing (FF) : identification of the most non-influential factors (screening)
- To make **diagnosis** on :
  - The **driving forces** of plant growth and development
  - The relative importance of the described **biophysical processes** regarding the outputs of interest

The independent case : Hoeffding decomposition (1948)

Assume that $(X_i)_{i\in\{1:p\}}$ are independent parameters and η a model.

Theorem (Functional decomposition of η)

We have the unique decomposition of the model η

$$\begin{aligned}\eta(X) &= \eta_0 + \sum_{i=1}^{p}\eta_i(X_i) + \sum_{i,j=1,i\neq j}^{p}\eta_{i,j}(X_{i,j}) + \cdots + \eta_{1,\ldots,p}(X) \\ &= \sum_{u\in\{1:p\}}\eta_u(X_u).\end{aligned} \tag{1}$$

where X_u is a group of variables, η_u only depends on X_u and

$$\int \eta_u(x_u)\eta_v(x_v)d\mathbb{P}_X = \mathbb{E}(\eta_u(X_u)\eta_v(X_v)) = 0, \quad \forall u, v \subseteq \{1:p\}, \quad u \neq v$$

ANOVA decomposition and Sobol's indexes (Sobol, 1993)

- Analysis of Variance (ANOVA) decomposition

$$\mathbb{V}(Y) = \sum_{u} \mathbb{V}(\eta_u(X_u)) = \sum_{i=1}^{p} V_i + \sum_{1 \le i < j \le p} V_{ij} + \cdots + V_{1,2,\ldots,p}$$

- Sobol's indexes

$$S_u = \frac{\mathbb{V}(\eta_u)}{\mathbb{V}(Y)} = \frac{\mathbb{V}(\mathbb{E}[Y|X_u]) - \sum_{v \subsetneq u} \mathbb{V}(\mathbb{E}[Y|X_v])}{\mathbb{V}(Y)}.$$

ANOVA decomposition and Sobol's indexes (Sobol, 1993)

- Analysis of Variance (ANOVA) decomposition

$$\mathbb{V}(Y) = \sum_{u} \mathbb{V}(\eta_u(X_u)) = \sum_{i=1}^{p} V_i + \sum_{1 \le i < j \le p} V_{ij} + \cdots + V_{1,2,\ldots,p}$$

- Sobol's indexes

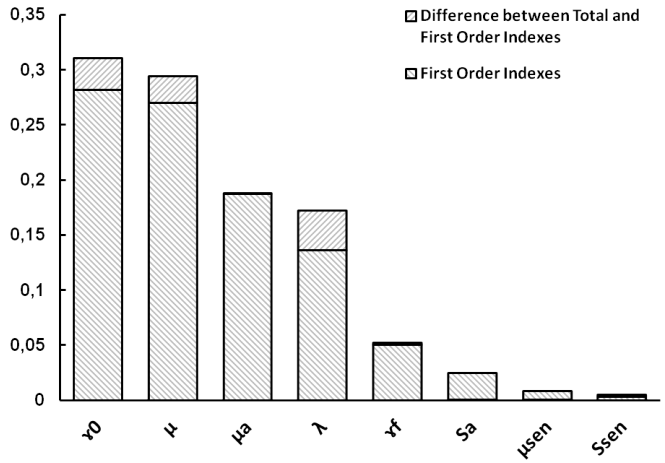
$$S_u = \frac{\mathbb{V}(\eta_u)}{\mathbb{V}(Y)} = \frac{\mathbb{V}(\mathbb{E}[Y|X_u]) - \sum_{v \subsetneq u} \mathbb{V}(\mathbb{E}[Y|X_v])}{\mathbb{V}(Y)}.$$

- First-order index : $S_i = \frac{V_i}{\mathbb{V}(Y)}$
  for '**Factor Priorization**'
- Total index : $S_i^T = S_i + \sum_{j \neq i} S_{i,j} + \sum_{j \neq i, k \neq i, j<k} S_{i,j,k} + \cdots + S_{1,\ldots,p}$.
  for '**Factor Fixing**'
- $1 = \sum_{i=1}^{p} S_i + \sum_{1 \le i < j \le p} S_{ij} + \cdots + S_{1,2,\ldots,p}$
  $\sum_{i=1}^{p} S_i$ serves as '**Model Linearity Index**

Sobol's Methods to support Parameter Estimation

SA analysis for the LNAS model [C. et al., 2013]

Output chosen : related to the criterion to optimize for parameter estimation.



$\Longrightarrow$ Help rank the parameters and then process parameter estimation with an increasing number of params. (the others being fixed to their nominal values)

Nb. of est. params.	1	2	3	4	5	6	7	8
AICc	351.5	346.9	346.0	347.2	343.0	346.0	347.8	348.8

Table : Corrected AIC for LNAS model with 1 to 8 estimated parameters

Exemple of Model Diagnosis

Non-linearity assessment : GreenLab Maize [Wu et al., 2009]

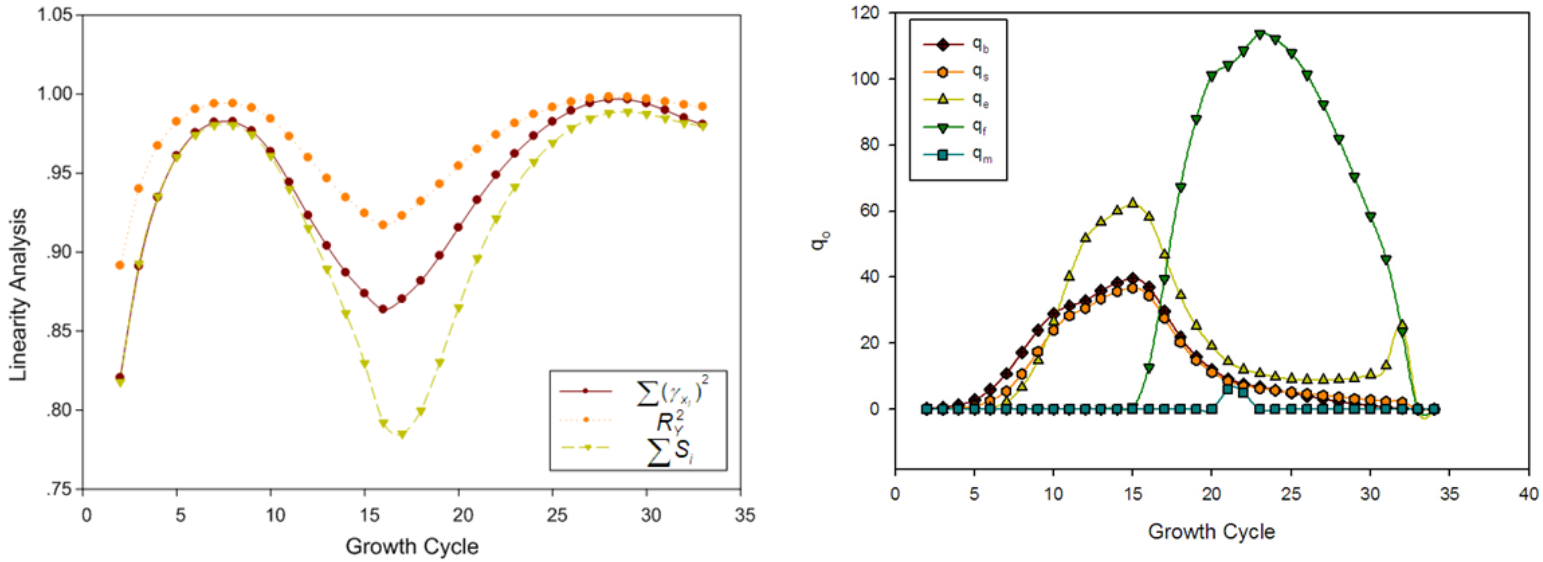


Figure : GreenLab Maize (a) Evolution of the linearity index with output of biomass production (b) At each GC, biomass allocation per organ type (b : leaf blade ; s : sheath ;e : internode ; f : cob ; m : tassel)

Exemple of Model Diagnosis

Non-linearity assessment : GreenLab Maize [Wu et al., 2009]

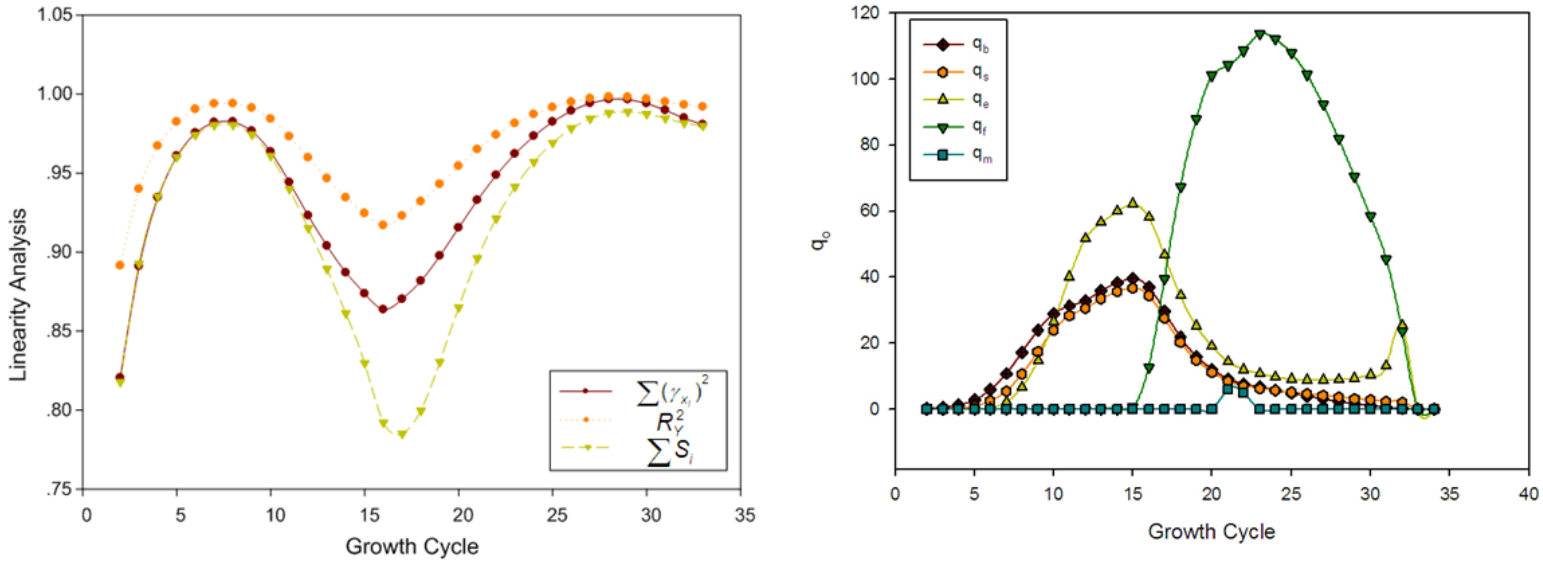


Figure : GreenLab Maize (a) Evolution of the linearity index with output of biomass production (b) At each GC, biomass allocation per organ type (b : leaf blade ; s : sheath ;e : internode ; f : cob ; m : tassel)

A non-linear period is denoted around GC17. $\Longrightarrow$ A key step in terms of **biophysical processes** corresponding to the **transition between two allocation phases**

Comprehensive Methodology for Complex Biophysical Systems [Wu et C., 2014]

Motivation :

- A complex biological system is characterized by several interacting processes with **submodels/modules** describing each of them

Comprehensive Methodology for Complex Biophysical Systems [Wu et C., 2014]

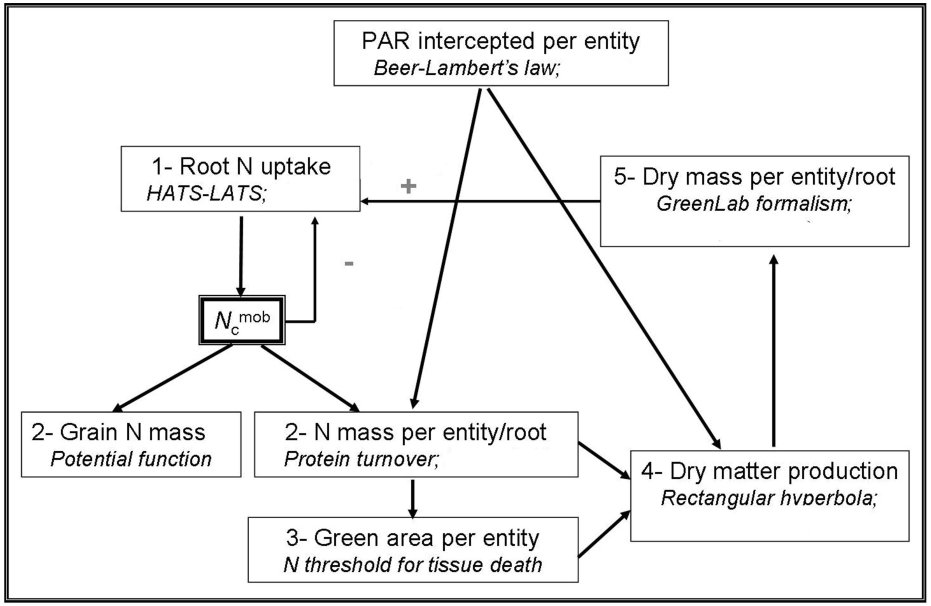
Motivation :

- A complex biological system is characterized by several interacting processes with **submodels/modules** describing each of them

Comprehensive Strategy

- **Step 1. Non-linearity study with SRC** : R^2
- **Step 2. Group analysis** : compute the sensitivity indices for each module and interactions between modules
- **Step 3. Internal module analysis** : screening most non-influential factors in each specific module
- **Step4. Overall model analysis with the selected parameters**

Application to NEMA model [Bertheloot, et al. 2011]



Five biological modules :

- Nitrogen acquisition by roots(**RootNuptake : 34 parameters**)
- Nitrogen distribution(**Nflux : 28 parameters**)
- Carbon acquisition via photosynthesis(**Photosynthesis : 10 parameters**)
- Carbon distribution(**DMflux : 5 parameters**)
- Senescence(**Tissuedeath : 5 parameters**)

$\Longrightarrow$ 17 influential parameters are identified (among 82) : drastic model simplification !

A difficulty : the dependent case

In plant growth models, there are usually correlations between parameters (due for example to pleiotropic genetic controls or correlated processes) !

$\Longrightarrow$ Sobol indexes are no longer relevant when the inputs X_i are dependent.

A difficulty : the dependent case

In plant growth models, there are usually correlations between parameters (due for example to pleiotropic genetic controls or correlated processes) !
$\Longrightarrow$ Sobol indexes are no longer relevant when the inputs X_i are dependent.

Example (Chastaing, 2012)

$$Y = \eta(X) = X_1 + X_2,$$
$$X \backsim \mathcal{N}(0, \Sigma) \text{ and } \Sigma = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}.$$

$$S_1 = \frac{(1+\sigma)^2}{2+2\sigma}, \quad S_2 = \frac{(1+\sigma)^2}{2+2\sigma}, \quad S_{12} = \frac{2\sigma^2(1+\sigma)}{2+2\sigma}.$$

Correlation	S_1	S_2	S_{12}	$\sum_u S_u$
$\sigma = 0$ (independent)	0.5	0.5	0	1
$\sigma = 0.9$ (dependent)	0.95	0.95	0.81	2.71

- In the dependent case, Hoeffding decomposition is not unique.
- $\sum_u S_u$ is not equal to 1. An information is taking into account several times.

ANCOVA (Li and Rabitz, 2010)

Vector spaces decomposition

Assume that $\eta \in H = L^2_{\mathbb{R}}(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_x)$ with the usual inner product $\langle f, g \rangle = \int f(x)g(x)dP_x$ for H.
$\forall u \in \{1:p\}$, H_u is the vector space of function only depending on X_u.
Let be the family of vector subspaces $(H^0_u)_{u \in S}$:

- $H^0_\emptyset = H_\emptyset$ is the set of constant functions
- and satisfying the hierachical orthogonality property

$$\forall u \in S^*, \quad H^0_u = \left\{ h_u \in H_u \mid \langle h_u, h_v \rangle = 0, \ \forall v \subset u, \ \forall h_v \in H^0_v \right\}. \tag{2}$$

Chastaing (2012) gives a uniqueness result for dependent inputs : generalization of Hoeffding decomposition.

$$H = \bigoplus_{u \in \{1:p\}} H^0_u \tag{3}$$

$$Y = \eta(X) = \sum_{u \in \{1:p\}} \eta_u(X_u) \tag{4}$$

ANCOVA (Li and Rabitz, 2010)

AnCoVa decomposition

$$
\begin{aligned}
\mathbb{V}(Y) &= Cov(Y, Y), \\
&= Cov\left(Y, \sum_{i=1}^{p} \eta_i(X_i) + \cdots + \eta_{1,\ldots,p}(X)\right), \\
&= \underbrace{\sum_{u \subset \{1:p\}} \mathbb{V}(\eta_u(X_u))}_{ANOVA} + \underbrace{\sum_{u \subset \{1:p\}} \sum_{\substack{v \subset \{1:p\} \\ u \cap v \neq \{u,v\}}} Cov(\eta_u(X_u), \eta_v(X_v))}_{\text{correlated terms}}.
\end{aligned}
\tag{2}
$$

ANCOVA (Li and Rabitz, 2010)

Generalized Sobol (gSobol) indexes

- Total contribution of X_u

$$gS_u = \frac{Cov\left(Y, \eta_u(X_u)\right)}{\mathbb{V}(Y)}$$

- Structural contribution of X_u

$$gS_u^S = \frac{\mathbb{V}\left(\eta_u(X_u)\right)}{\mathbb{V}(Y)}$$

- Correlative contribution of X_u

$$gS_u^C = \frac{1}{\mathbb{V}(Y)} \sum_{\substack{v \subset \{1:p\} \\ u \cap v \neq \{u,v\}}} Cov(\eta_u(X_u), \eta_v(X_v)).$$

We have $S_u = S_u^S + S_u^C$.

$\Longrightarrow$ The conclusions can be very different !

Application to LNAS : 10 parameters, Output : Dry Green Leaf Mass [Sainte-Marie et al., 2016]

time interval : $[80, 160]$ - time step : 5 jours. 2 independent groups of parameters :

- a first group involved in foliar senescence dynamics : $(tt_{sen}, mu_{sen}, s_{sen})^\top \sim \mathcal{N}_3(\mu_3, \Sigma_3)$ where $\mu_3 = (644;\ 2400;\ 4520)^\top$ and $\Sigma_3 = \sigma_3 \rho_3 \sigma_3$ with $\sigma_3 = (32,2;\ 120;\ 226)^\top$ and

$$\rho_3 = \begin{bmatrix} 1 & 0,5 & -0,5 \\ 0,5 & 1 & 0,2 \\ -0,5 & 0,2 & 1 \end{bmatrix}.$$

- a second group involved in allocation dynamics between roots and leaves : $(mu_{alloc}, s_{alloc}, s_{init}, s_{end})^\top \sim \mathcal{N}_4(\mu_4, \Sigma_4)$ where $\mu_4 = (550;\ 300;\ 0,7;\ 0,15)^\top$ and $\Sigma_4 = \sigma_4 \rho_4 \sigma_4$ with $\sigma_4 = (27,5;\ 15;\ 0,035;\ 0,075)^\top$ and

$$\rho_4 = \begin{bmatrix} 1 & 0,2 & 0 & 0 \\ 0,2 & 1 & 0,5 & -0,5 \\ 0 & 0,5 & 1 & -0,5 \\ 0 & -0,5 & -0,5 & 1 \end{bmatrix}.$$

3 additional independent parameters $rue \sim \mathcal{N}(3,6;\ 0,15)$, $e \sim \mathcal{N}(60;\ 3)$, $k_b \sim \mathcal{N}(0,7;\ 0,035)$

Application to LNAS : 10 parameters, Output : Dry Green Leaf Mass [Sainte-Marie et al., 2016]

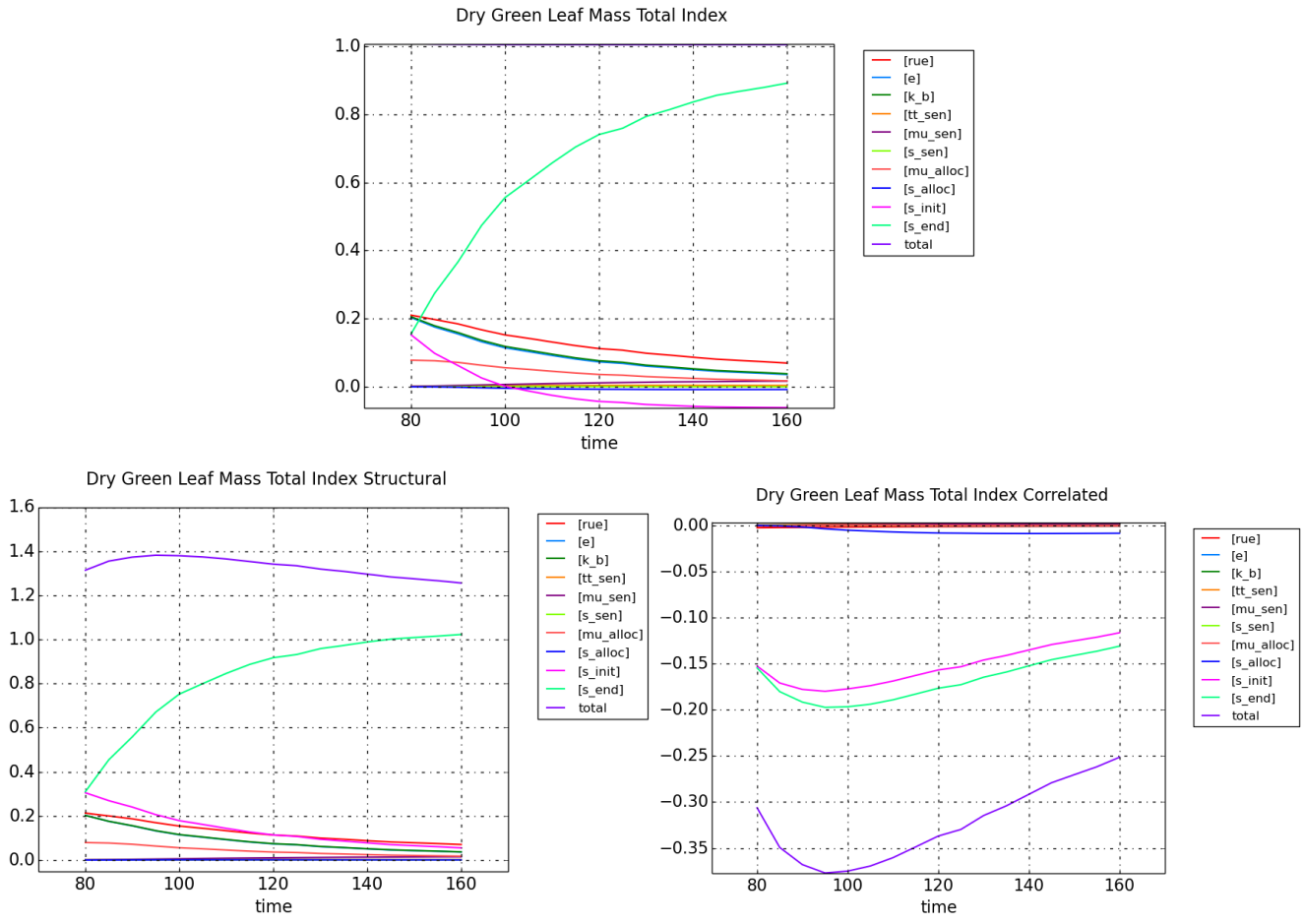


Figure : Generalized Sobol indexes associated to the Dry Green Leaf Mass

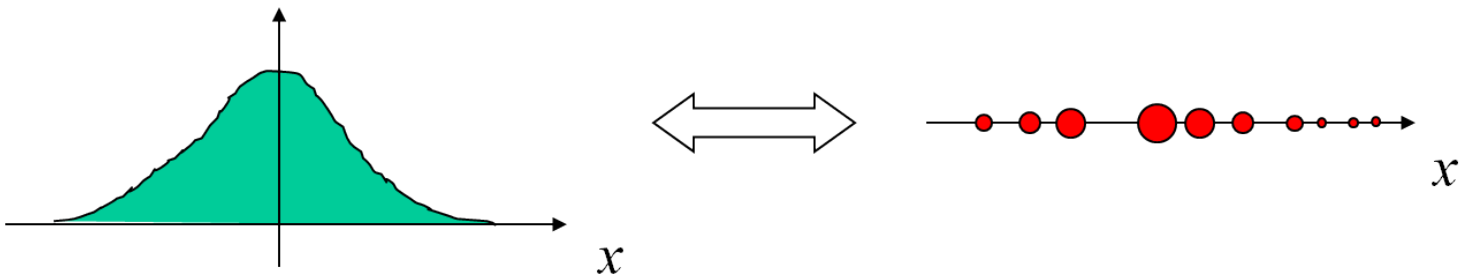
Outline

Fundamental concepts of Sequential Monte Carlo methods

- Objective of Bayesian filtering methods :
  provide an estimator $\hat{p}(x_n^a|y_{1:n})$ of $p(x_n^a|y_{1:n})$, where $x_n^a = (\theta, x_n)$.
- Monte Carlo Samples (Particles)

Fundamental concepts of Sequential Monte Carlo methods

- Objective of Bayesian filtering methods :
  provide an estimator $\hat{p}(x_n^a|y_{1:n})$ of $p(x_n^a|y_{1:n})$, where $x_n^a = (\theta, x_n)$.
- Monte Carlo Samples (Particles)



$\Rightarrow$ A value and a weight assigned to each particle

$\Rightarrow$ ideal case : drawn directly from $p(x_n^a|y_{0:n})$ (too difficult)

$\Rightarrow$ in practice : drawn from (importance sampling)

Algorithm for Convolution Particle Filtering

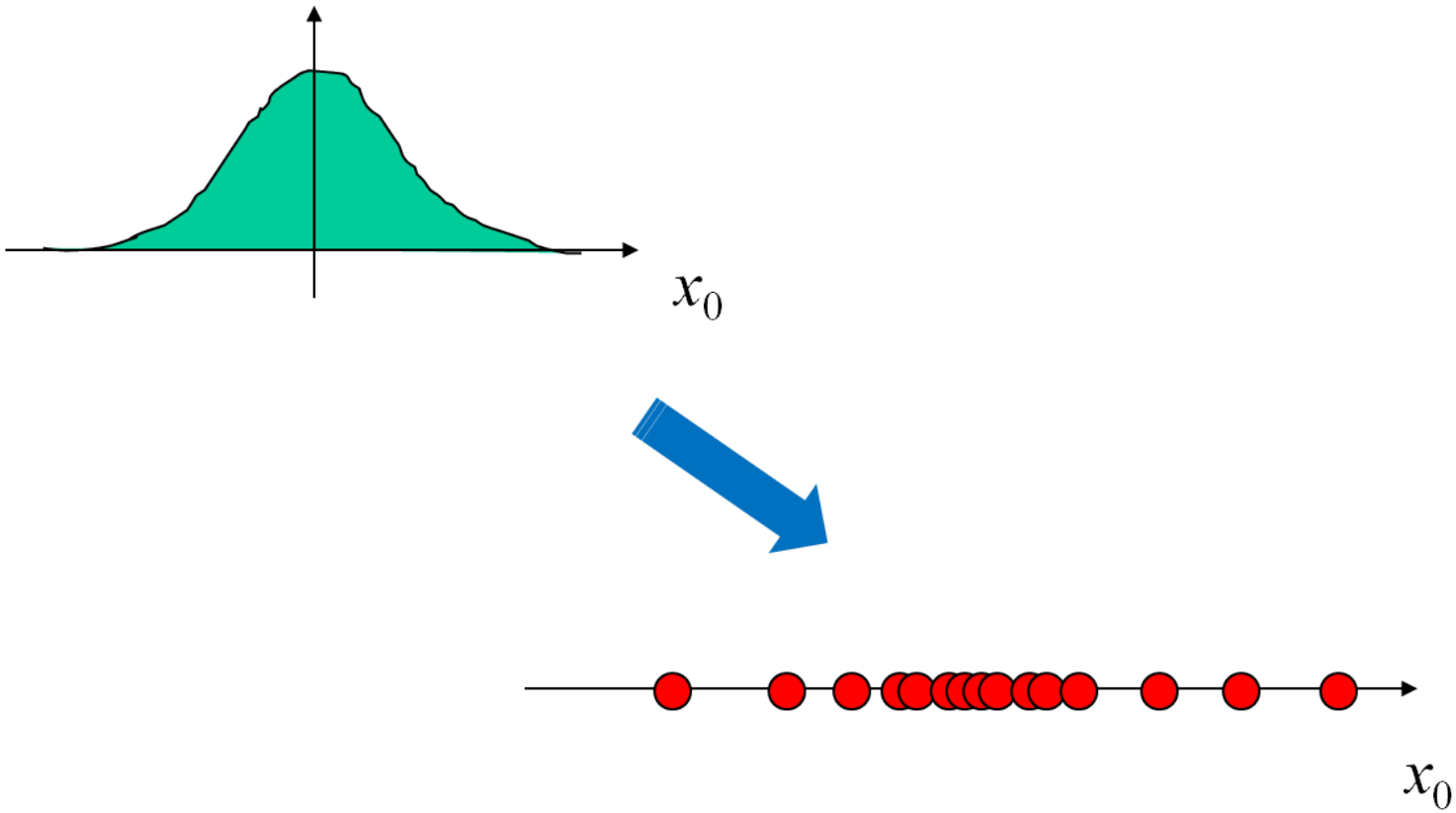
[Campillo *et al.*, 2009]

- **Initialization** of the particles. For $i = 1, \ldots, M$, $\tilde{x}_0^{a(i)} \sim p(x_0^a)$, $w_0(\tilde{x}_0^{a(i)}) = \frac{1}{M}$

Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Initialization** of the particles. For $i = 1, \ldots, M$, $\tilde{x}_0^{a(i)} \sim p(x_0^a)$, $w_0(\tilde{x}_0^{a(i)}) = \frac{1}{M}$



Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Initialization** of the particles. For $i = 1, \ldots, M$, $\tilde{x}_0^{a(i)} \sim p(x_0^a)$, $w_0(\tilde{x}_0^{a(i)}) = \frac{1}{M}$
- **Iteration** For $n = 0, \ldots, N$,

  $ **Prediction**

  $ **Correction**

Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Initialization** of the particles. For $i = 1, \ldots, M$, $\tilde{x}_0^{a(i)} \sim p(x_0^a)$, $w_0(\tilde{x}_0^{a(i)}) = \frac{1}{M}$
- **Iteration** For $n = 0, \ldots, N$,
  - $ **Prediction** : For $i = 1, \ldots, M$, $\tilde{x}^a_{n+1-}{}^{(i)} \sim p(x^a_{n+1}|\tilde{x}^{a(i)}_n)$, $\tilde{y}^{(i)}_{n+1-} \sim p(y_{n+1}|\tilde{x}^a_{n+1-}{}^{(i)})$

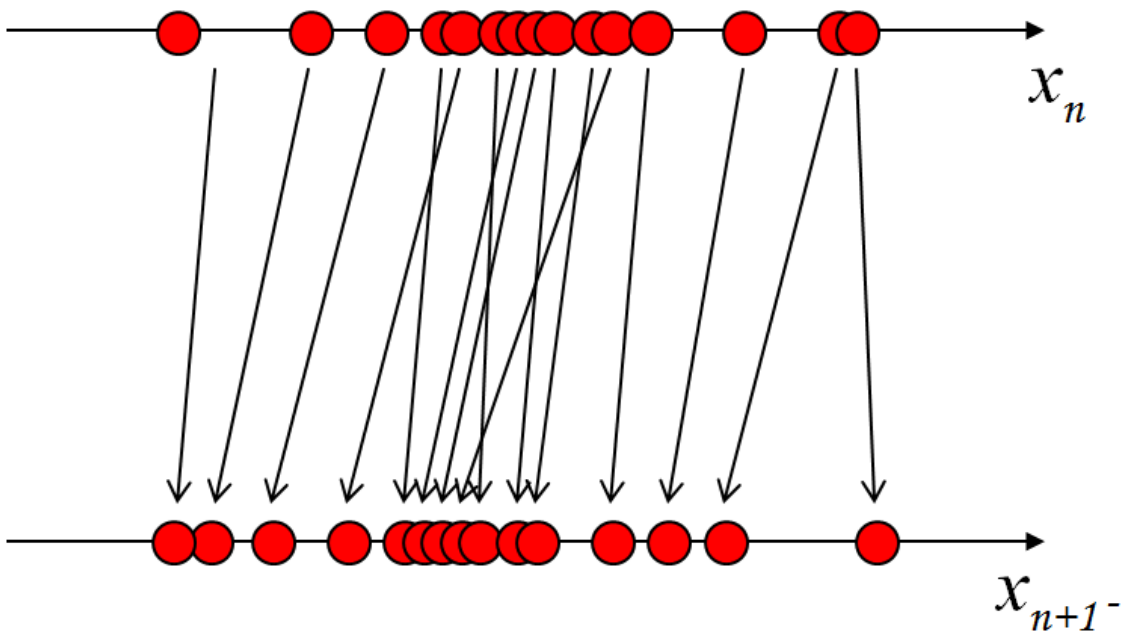
    Weight calculation :

$$w^{(i)}_{n+1} = K^Y_{h_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1-})$$

Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Iteration** For $n = 0, \ldots, N$,
  - $ **Prediction** : For $i = 1, \ldots, M$,
    $\tilde{x}^a_{n+1-}{}^{(i)} \sim p(x^a_{n+1}|\tilde{x}^{a(i)}_n)$, $\tilde{y}^{(i)}_{n+1-} \sim p(y_{n+1}|\tilde{x}^a_{n+1-}{}^{(i)})$

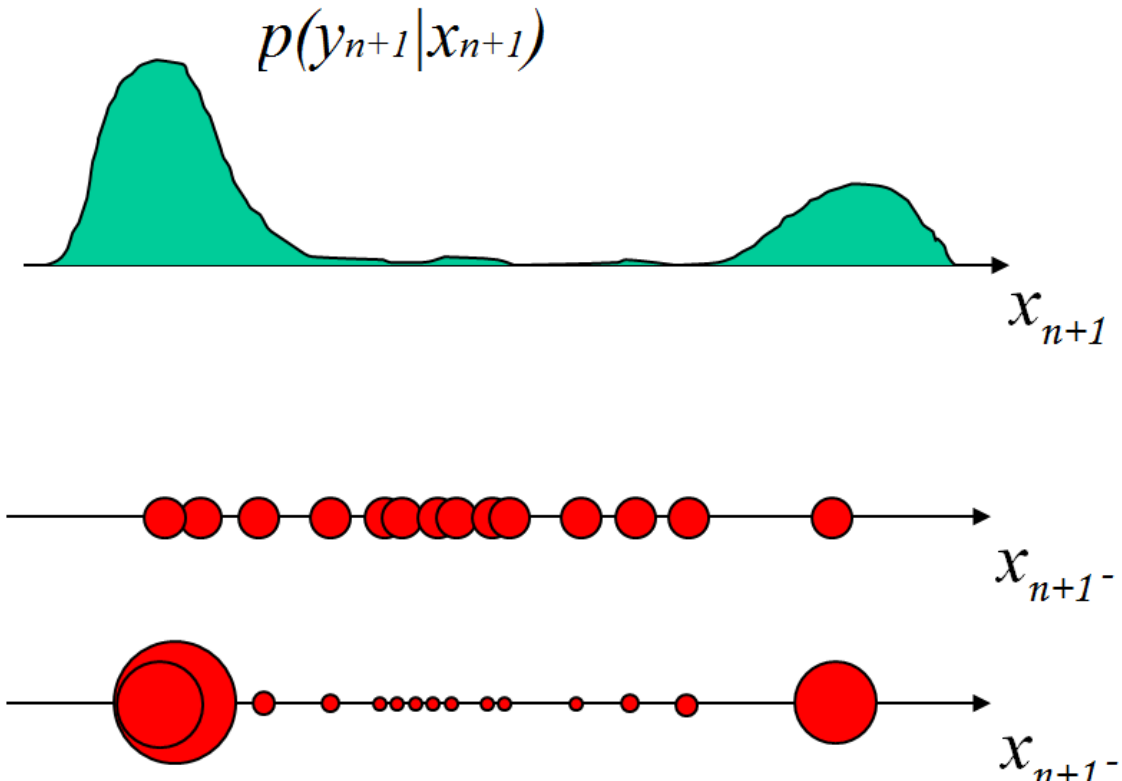


Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Iteration** For $n = 0, \ldots, N$,
  - $ **Prediction** : For $i = 1, \ldots, M$,
    $\tilde{x}^a_{n+1-}{}^{(i)} \sim p(x^a_{n+1}|\tilde{x}^{a(i)}_n)$, $\tilde{y}^{(i)}_{n+1-} \sim p(y_{n+1}|\tilde{x}^a_{n+1-}{}^{(i)})$
    Weight calculation :

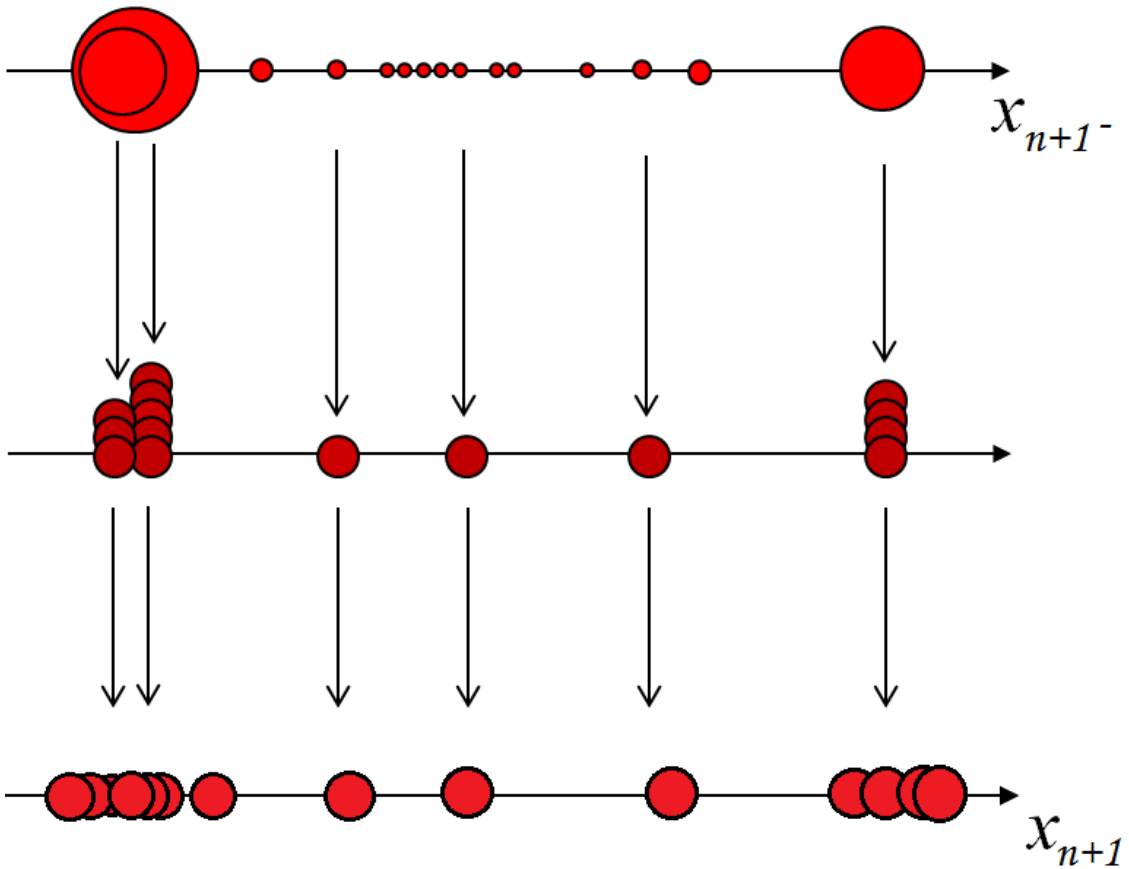
$$w^{(i)}_{n+1} = K^Y_{h_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1-})$$



Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Iteration** For $n = 0, \ldots, N$,
  - $ **Correction** : For $i = 1, \ldots, M$,
    $\tilde{x}^a_{n+1}{}^{(i)} \sim \hat{p}(x^a_{n+1}|y_{0:n+1})$



Algorithm for Convolution Particle Filtering

[Campillo *et al.*, 2009]

- **Initialization** of the particles. For $i = 1, \ldots, M$, $\tilde{x}_0^{a(i)} \sim p(x_0^a)$, $w_0(\tilde{x}_0^{a(i)}) = \frac{1}{M}$
- **Iteration** For $n = 0, \ldots, N$,
  - $ **Prediction** : For $i = 1, \ldots, M$, $\tilde{x}^a_{n+1-}{}^{(i)} \sim p(x^a_{n+1}|\tilde{x}^{a(i)}_n)$, $\tilde{y}^{(i)}_{n+1-} \sim p(y_{n+1}|\tilde{x}^a_{n+1-}{}^{(i)})$

    Weight calculation :

    $$w^{(i)}_{n+1} = K^Y_{h_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1-})$$

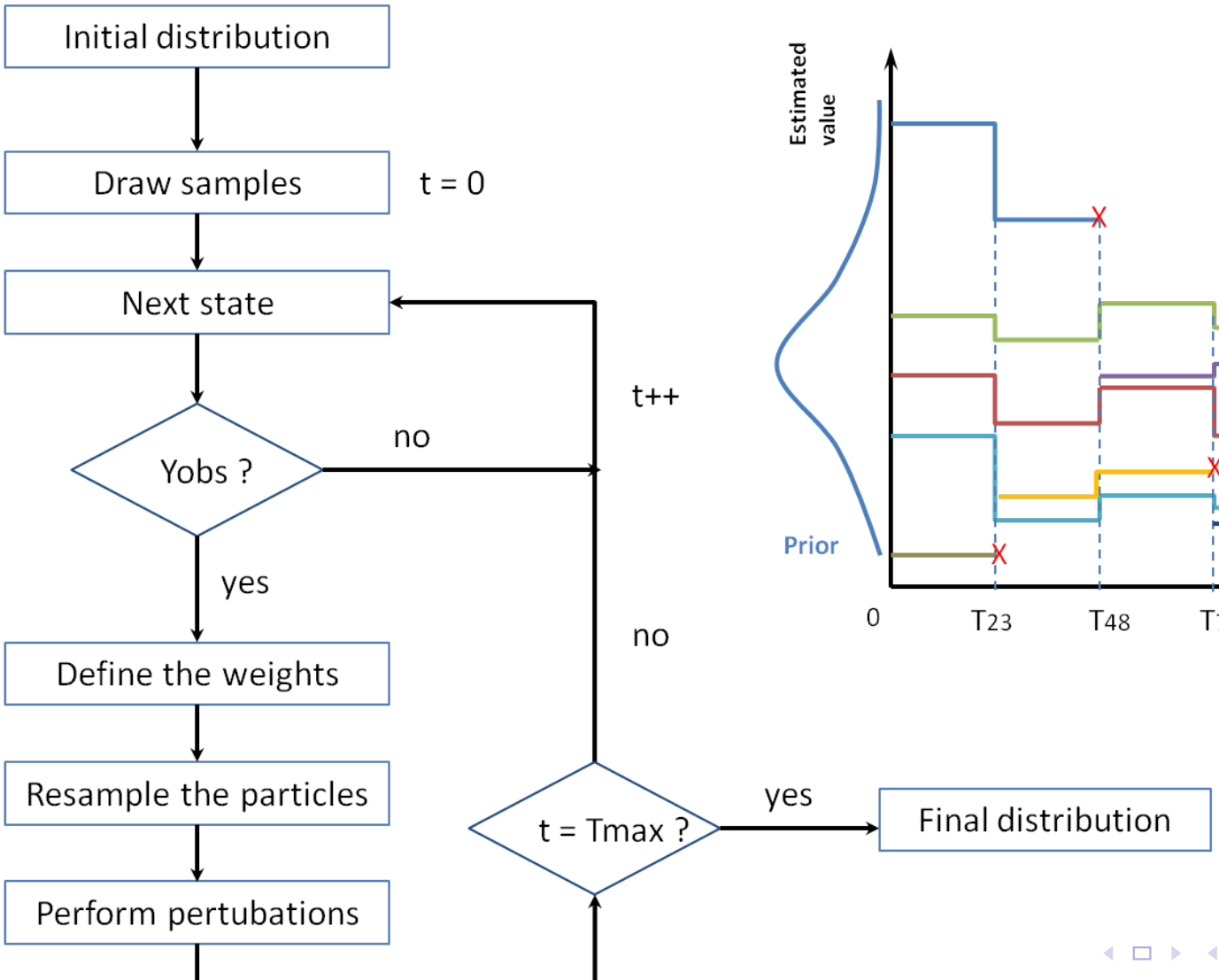
  - $ **Correction** : For $i = 1, \ldots, M$, $\tilde{x}^a_{n+1}{}^{(i)} \sim \hat{p}(x^a_{n+1}|y_{0:n+1})$

    Kernel based estimator :

    $$\hat{p}(x^a_{n+1}|y_{0:n+1}) = \sum_{i=1}^{M} \tilde{w}^{(i)}_{n+1} K^X_{h_M}(x^a_{n+1} - \tilde{x}^a_{n+1-}{}^{(i)})$$

Algorithm for Convolution Particle Filtering

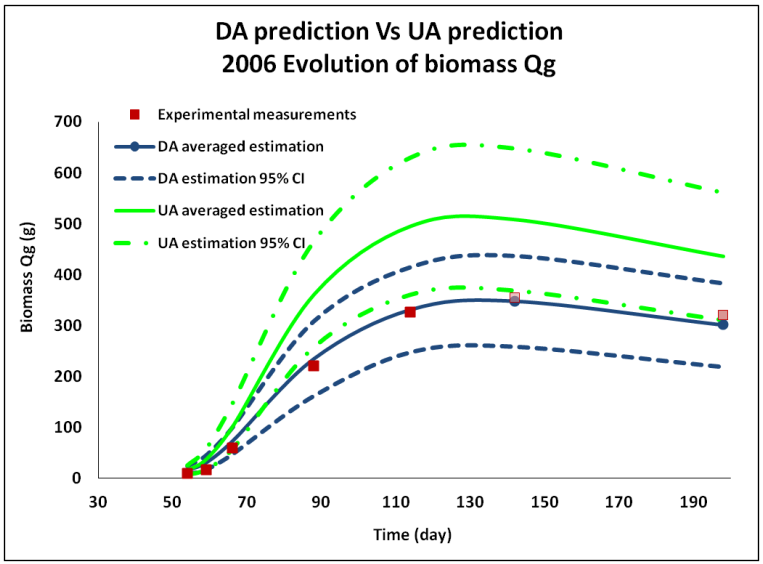
[Campillo *et al.*, 2009]



Data Assimilation for Prediction

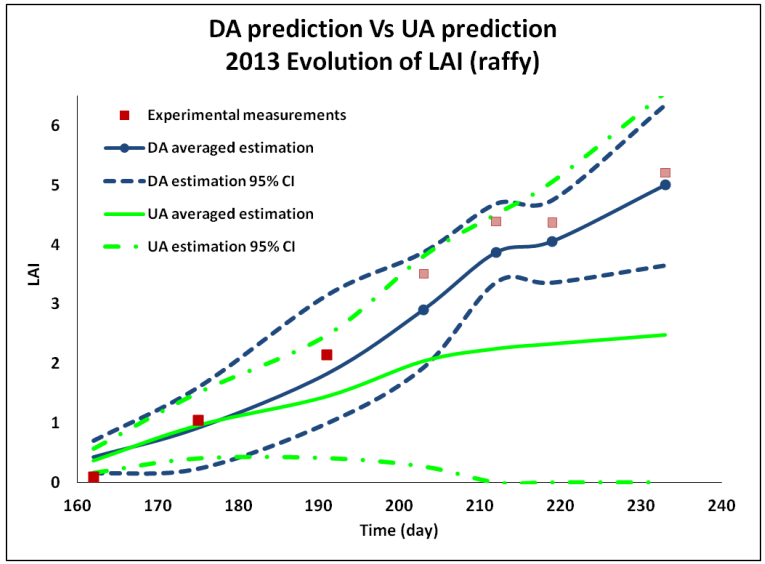
$\Longrightarrow$ Use of experimental data in the early stages of crop growth to estimate $p\left(\theta, x_n | y_{\leq n}\right)$ and then predict the final $p\left(x_N\right)$

- Sugar beet with LNAS [Chen, 2014]



	Real Data 2006	DA estimates (relative error in %)	95% CI	UA estimates (relative error in %)	95% CI
$Q_b\,(t_{142})$	355.2	348.1 (2.0%)	[258.7; 437.4]	507.8 (43.0%)	[368.4; 647.3]
$Q_b\,(t_{198})$	320.6	301.3 (6.0%)	[219.0; 383.6]	435.7 (35.9%)	[384.3; 560.7] *
$Q_r\,(t_{142})$	1459.2	1716.2 (17.6%)	[1427.9; 2004.5]	1930.7 (32.3%)	[1603.0; 2258.4] *
$Q_r\,(t_{198})$	2400.0	2644.3 (10.2%)	[2209.4; 3079.2]	2942.9 (22.6%)	[2455.0; 3430.7] *

- Wheat with STICS [Chen, 2014]



	Real Data 2013	DA estimates (relative error in %)	95% CI	UA estimates (relative error in %)	95% CI
$hur\,(t_{198})$	0.276	0.272 (1.7%)	[0.260; 0.283]	0.290 (5.0%)	[0.286; 0.295]
$hur\,(t_{210})$	0.274	0.303 (10.3%)	[0.284; 0.322]	0.308 (12.2%)	[0.303; 0.313]
$hur\,(t_{226})$	0.289	0.297 (2.6%)	[0.278; 0.316]	0.303 (4.7%)	[0.295; 0.310]
$hur\,(t_{255})$	0.239	0.262 (9.7%)	[0.240; 0.283]	0.272 (14.0%)	[0.253; 0.291]
$hur\,(t_{275})$	0.257	0.260 (1.2%)	[0.229; 0.291]	0.267 (4.0%)	[0.233; 0.302]
$LAI\,(t_{203})$	4.235	4.514 (6.6%)	[3.144; 5.883]	3.482 (17.8%)	[0.471; 6.494]
$LAI\,(t_{212})$	4.735	4.627 (2.3%)	[3.192; 6.062]	3.842 (18.9%)	[0.000; 7.693]
$LAI\,(t_{219})$	**4.225**	4.867 (15.2%)	[2.394; 7.340]	3.994 (5.5%)	[0.000; 8.638]
$LAI\,(t_{233})$	4.910	5.215 (6.2%)	[0.969; 9.462]	4.274 (13.0%)	[0.000; 11.214]
$magrain\,(t_{266})$	819.38	804.33 (1.8%)	[611.68; 996.98]	550.58 (32.8%)	[128.97; 972.20]

Outline

Statistical Framework

A population of Genotypes ...

- Typical situation : a small number of plants are measured for a family of genotypes
- The genetic variability will be studied with a population-based model with the genotype as the random effect [Baey et al., 2014]

Statistical Framework

A population of Genotypes ...

- Typical situation : a small number of plants are measured for a family of genotypes
- The genetic variability will be studied with a population-based model with the genotype as the random effect [Baey et al., 2014]

... Represented by a Hierarchical Mixed-Effect Model

- **First-stage :** intra-genotypic variation (for each genotype i, of param ϕ_i)

$$\begin{aligned} y_i &= F(\phi_i, x_i) + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}(0, \Sigma), \end{aligned}$$

- **Second-stage :** inter-genotypic variation

$$\begin{aligned} \phi_i &= \beta + \xi_i, \\ \xi_i &\sim \mathcal{N}_P(0, \Gamma), \end{aligned}$$

Statistical Framework

A population of Genotypes ...

- Typical situation : a small number of plants are measured for a family of genotypes
- The genetic variability will be studied with a population-based model with the genotype as the random effect [Baey et al., 2014]

... Represented by a Hierarchical Mixed-Effect Model

- **First-stage :** intra-genotypic variation (for each genotype i, of param ϕ_i)

$$\begin{aligned} y_i &= F(\phi_i, x_i) + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}(0, \Sigma), \end{aligned}$$

- **Second-stage :** inter-genotypic variation

$$\begin{aligned} \phi_i &= \beta + \xi_i, \\ \xi_i &\sim \mathcal{N}_P(0, \Gamma), \end{aligned}$$

- ▷ The genotypic variability is represented by the variability of the random parameters, i.e. by the covariance matrix Γ.
- ▷ We test for the variability of parameters by testing if their variances are null

Maximum Likelihood estimation for Mixed Models

- Parameters : $\theta = (\beta, \Gamma, \sigma^2), \theta \in \mathbb{R}^m$
- Likelihood :

$$L(\theta) := f(y;\theta) = \int_{\mathbb{R}^{p\times N}} f(y,\phi;\theta)d\phi = \int_{\mathbb{R}^{p\times N}} f(y \mid \phi;\theta)f(\phi;\theta)d\phi$$

- The nonlinearity of $g(t_{ij},\phi_i) = \mathbb{E}(y_{ij}|\phi_i)$ generally makes the computation of this integral untractable analytically
- Mixed models as incomplete data problem by considering random effects as missing data.

$\Rightarrow$ stochastic variants of Expectation- Maximization (EM) algorithm.

EM Algorithm

The main idea is to work with the density of the complete data $f(y, \phi; \theta)$.At iteration k :

- **Step E** (Expectation, with MCMC) : our objective is to approximate

$$Q(\theta; \theta^k) = \mathbb{E}\left(\log f(y, \phi; \theta) \mid y; \theta^k\right).$$

based on the generation of a Markov chain $(\phi^{k,(1)}, \ldots, \phi^{k,(m_k)})$:

$$\hat{Q}(\theta; \theta^k) = \frac{1}{m_k} \sum_{m=1}^{m_k} \log f(y, \phi^{k,(m)}; \theta)$$

or when reusing previous simulations Stochastic Approximation [Kuhn and Lavielle, 2005] :

$$\hat{Q}(\theta; \theta^k) = \hat{Q}(\theta; \theta^{k-1}) + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} \log f(y, \phi^{k,(m)}; \theta) - \hat{Q}(\theta; \theta^{k-1})\right]$$

- **Step M** (Maximization) :

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta; \theta^k).$$

Variance Components Testing

We consider here a diagonal variance structure Γ for the mixed effects :

$$\Gamma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & (0) \\ & & \ddots & \\ (0) & & & \sigma_p^2 \end{pmatrix}$$

Variance Components Testing

We consider here a diagonal variance structure Γ for the mixed effects :

$$\Gamma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & (0) \\ & & \ddots & \\ (0) & & & \sigma_p^2 \end{pmatrix}$$

We consider tests of the form :

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1$$

with

$$\Theta_0 = \{0\}^q \times [0; +\infty)^{p-q} \times \Omega, \quad \Theta_1 = [0; +\infty)^p \times \Omega$$

$\Longrightarrow$ When testing if q variance components are null, we are thus testing if these q components are on the boundary of the parameter space Θ.

Case of one variance (variability of parameter k of mean β_k and variance σ_k^2)

$$H_0 : \{\sigma_k^2 = 0\} \quad \text{vs.} \quad H_1 : \{\sigma_k^2 \geq 0\},$$

Likelihood ratio test : $T = -2(\ell_0(\theta) - \ell_1(\theta)) \overset{H_0}{\sim} \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2$

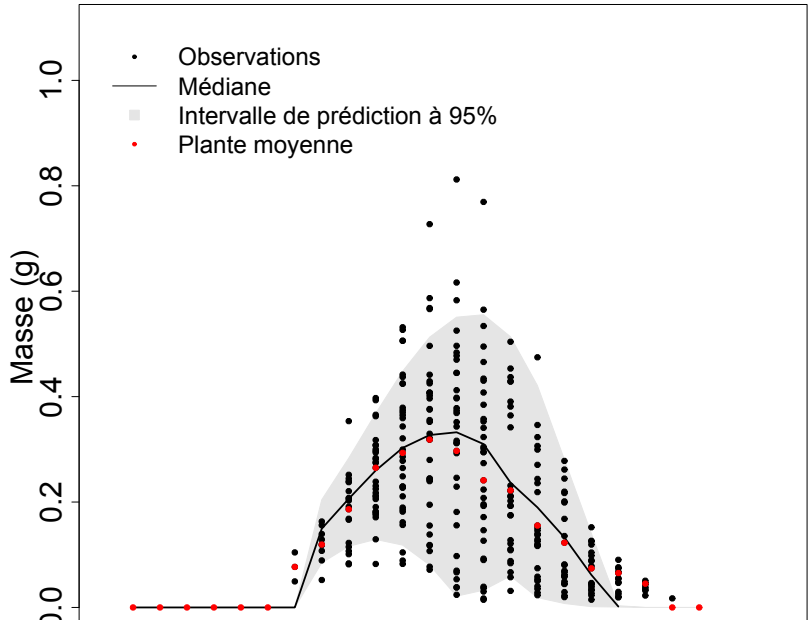
Application to the GreenLab model of Rapeseed

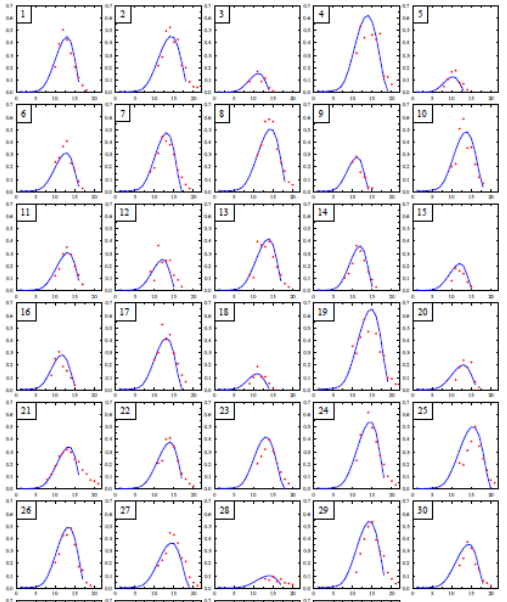
Collaboration with INRA Grignon [Baey et al., 2016]

- 34 individual plants ; "rosette" stage, leaf profiles
- 4 parameters : μ, s^{pr}, a_l, b_l
- MCMC-EM : Adaptive Metropolis with Global Scaling [Andrieu 2008]
- test random vs fixed effects (with Likelihood ratio tests)

Results

- μ, a_l variable in the population : 2 constant parameters b_l, s^{pr}





Outline

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation
- Sensitivity Analysis (especially Sobol's method) can help
  - to reduce the complexity of model parameterization
  - to provide insights about the model

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation
- Sensitivity Analysis (especially Sobol's method) can help
  - to reduce the complexity of model parameterization
  - to provide insights about the model
  - but be careful if input parameters are dependent ! ! !

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation
- Sensitivity Analysis (especially Sobol's method) can help
  - to reduce the complexity of model parameterization
  - to provide insights about the model
  - but be careful if input parameters are dependent ! ! !
- Data Assimilation ('online model re-calibration') can help reduce prediction uncertainty

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation
- Sensitivity Analysis (especially Sobol's method) can help
  - to reduce the complexity of model parameterization
  - to provide insights about the model
  - but be careful if input parameters are dependent ! ! !
- Data Assimilation ('online model re-calibration') can help reduce prediction uncertainty
- Mixed-effect plant growth models can be used to identify inter-genotype parameter variability, but 2 major difficulties :
  - parameter estimation of nonlinear mixed-effect models
  - statistical tests on variance components

Summary

- Importance of a proper assessment of parameter uncertainty for prediction and genotype differentiation
- Sensitivity Analysis (especially Sobol's method) can help
  - to reduce the complexity of model parameterization
  - to provide insights about the model
  - but be careful if input parameters are dependent ! ! !
- Data Assimilation ('online model re-calibration') can help reduce prediction uncertainty
- Mixed-effect plant growth models can be used to identify inter-genotype parameter variability, but 2 major difficulties :
  - parameter estimation of nonlinear mixed-effect models
  - statistical tests on variance components
- All the methods are implemented in a generic way in the PYGMALION platform at CentraleSupélec, with the recent possibility of connecting to external simulators (test cases with simulators in GroIMP, R ...)

THANKS !