Variational vs ensemble strategies to optimize the parameters of a terrestrial model

Philippe Peylin, Ernest Koffi, Diego Santaren, Frédéric Chevallier, Sylvain Kuppel

Laboratoire des Sciences du Climat et de l'Environnement

Gif sur Yvette, France

Objectives

- Application of Data – Assimilation techniques for land carbon cycle modeling
- Investigate the information content of various observation streams
- Which method is best appropriated ? (Variational, Monte-carlo)

Outline

- Context : Challenge with the C cycle
- Use of Flux-data to optimize land surface models
- Results of the model optimizations
- Performance of Variational vs Monte-Carlo ensemble optimizations
- Summary

Fate of Anthropogenic CO_2 Emissions (2000-2008)

1.4 PgC y^{-1}

+

7.7 PgC y^{-1}

4.1 PgC y^{-1}

45%

3.0 PgC y^{-1}

29%

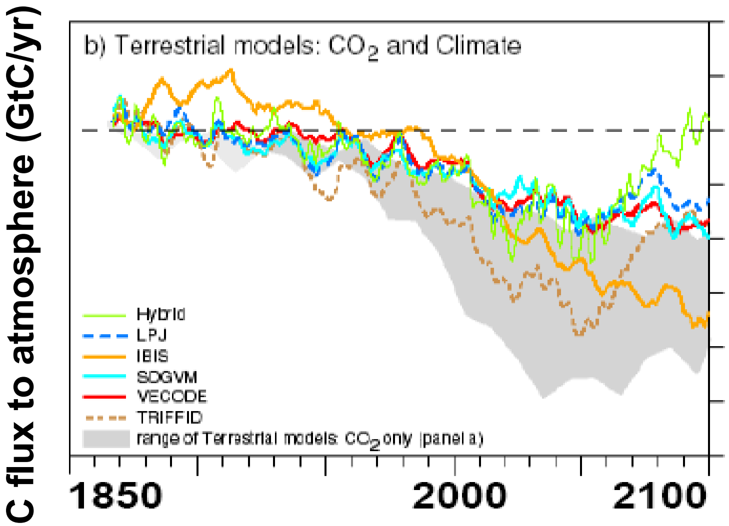
26%

2.3 PgC y^{-1}

Le Quéré et al. 2009, Nature-geoscience; Canadell et al. 2007, PNAS, updated

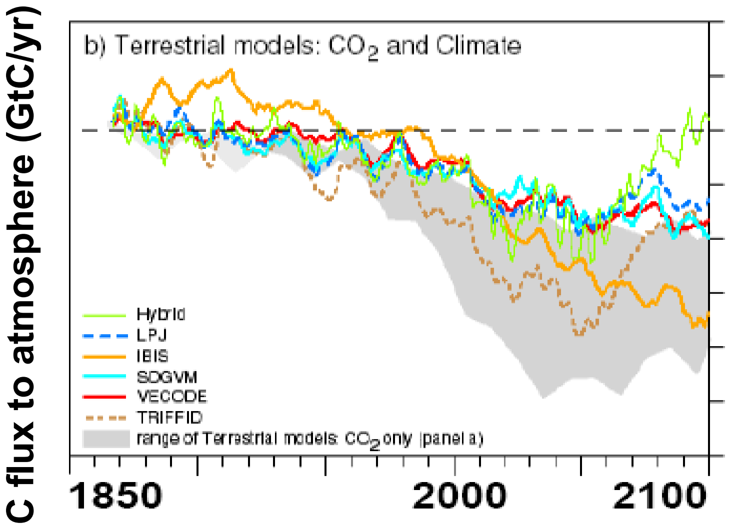
Needs for a Carbon Cycle Data Assimilation System

Large uncertainty from land to predict global C-balance (C4MIP)

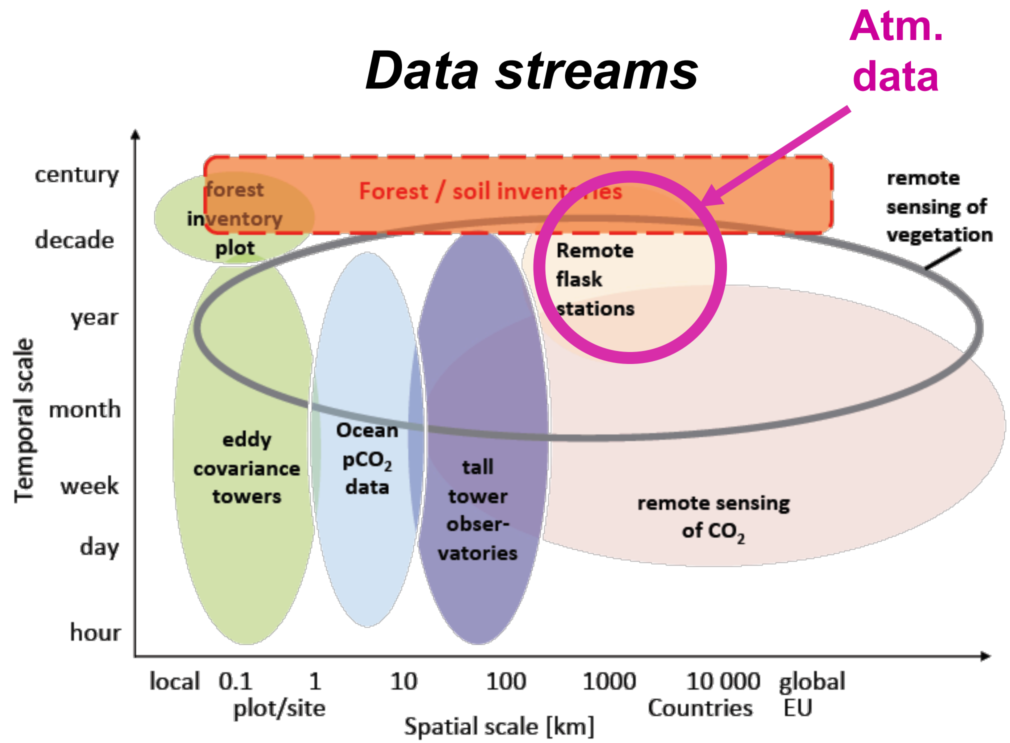


Needs for a Carbon Cycle Data Assimilation System

Large uncertainty from land to predict global C-balance (C4MIP)

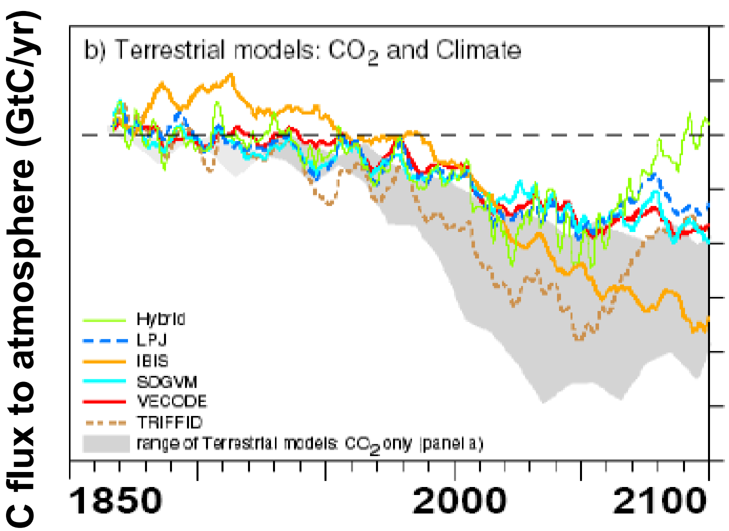


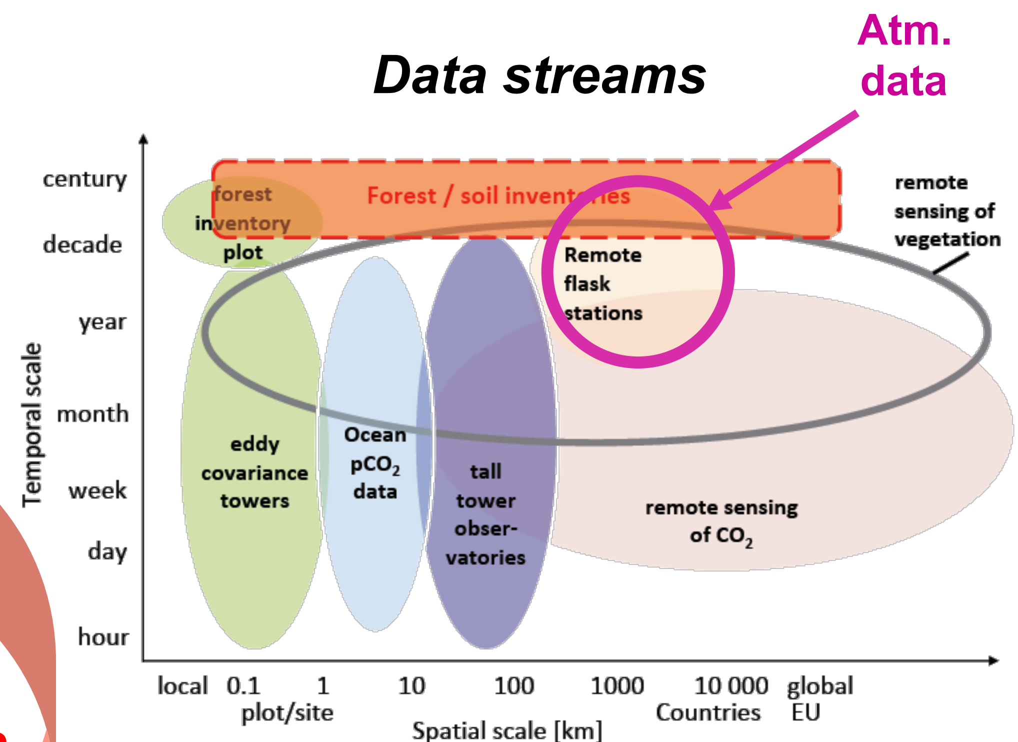
Data streams



Needs for a Carbon Cycle Data Assimilation System

Large uncertainty from land to predict global C-balance (C4MIP)



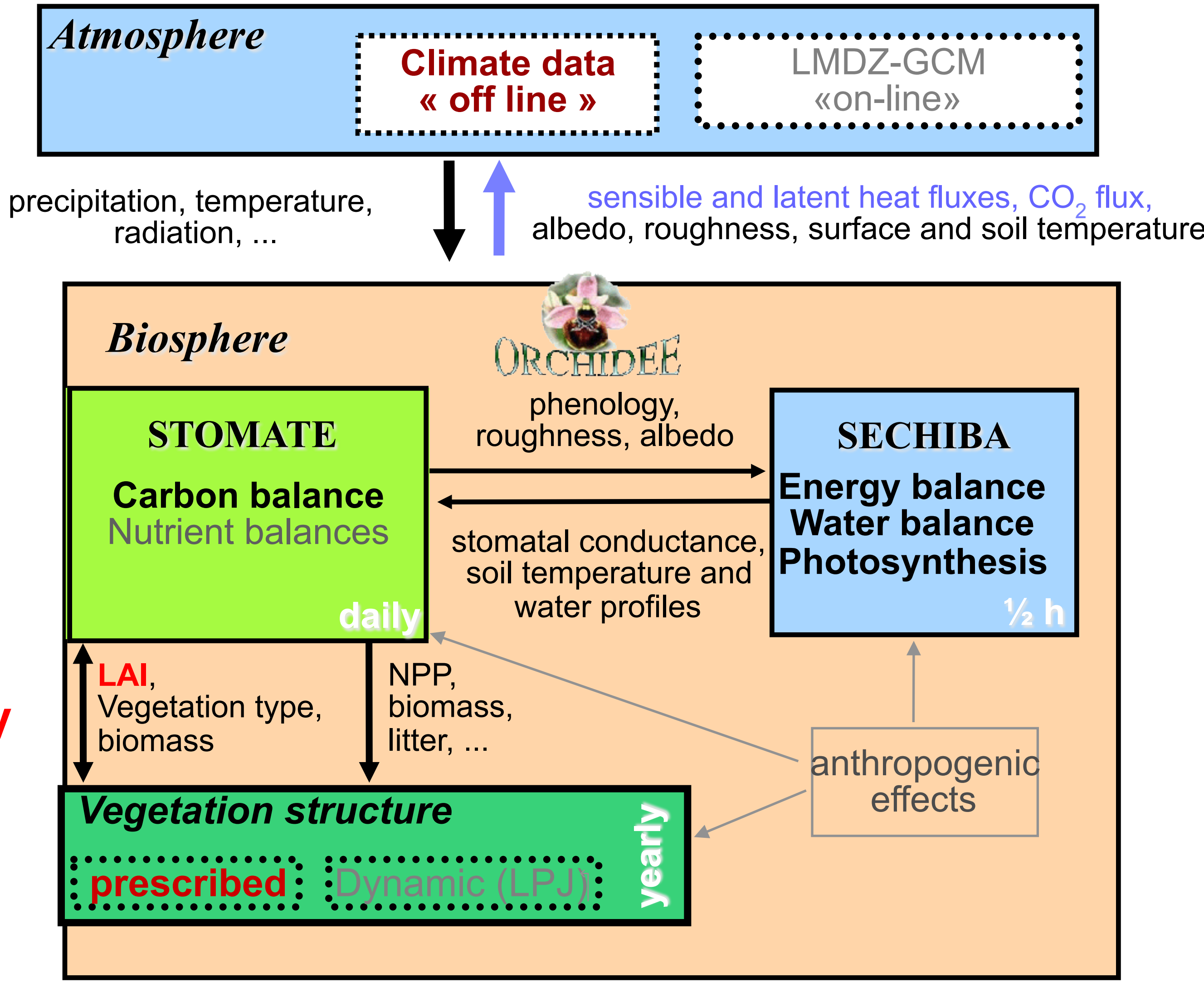


Data Assimilation

Optimized ecosystem models
➔ reduce the spread ?

The ORCHIDEE ecosystem model

- Process driven model used for IPCC AR5 simulations
- Energy / Water / Carbon balances
- Global - Site level
- 13 PFT's
- Pronostic phenology
- ½ hourly time step
- multiple C pools



Governing processes & parameters to optimize

Net SW downwelling Radiation

$$Rn = (1 - kalbedo)\, R^{i}_{lw} + R^{i}_{sw} - \varepsilon\sigma\, T^{4}_{surface}$$

Assimilation

$$A = Vc.(1 - \Gamma^{*}/C_i) - Rd \qquad Vc = \frac{fstress.Vc\ max\ .Ci}{Ci + Kc\ .(1 + \frac{O_i}{K_0})}$$

Evapo transpiration

$$A = 1/r_a + r_s\, (C_a - C_i)$$

$$ET = \sum_i \frac{K_E}{r_a + r_s} (q_i - q_{air})$$

$$1/r_s = gslope\ A.RH/C_a + g_0$$

Sensible heat

$$H = \frac{kcapa\ C_{sol}}{r_a} (T_{surf} - T_{air})$$

$$r_a = \frac{1}{kra.C_d(krugo)}$$

Maintenance resp.

Growth resp.

$$Rm = krespm \sum_i \lambda_i B_i\, f(T_{surface})$$

Heterotrophic resp.

$$Rg = krespg\, (A - Rm)$$

NEE

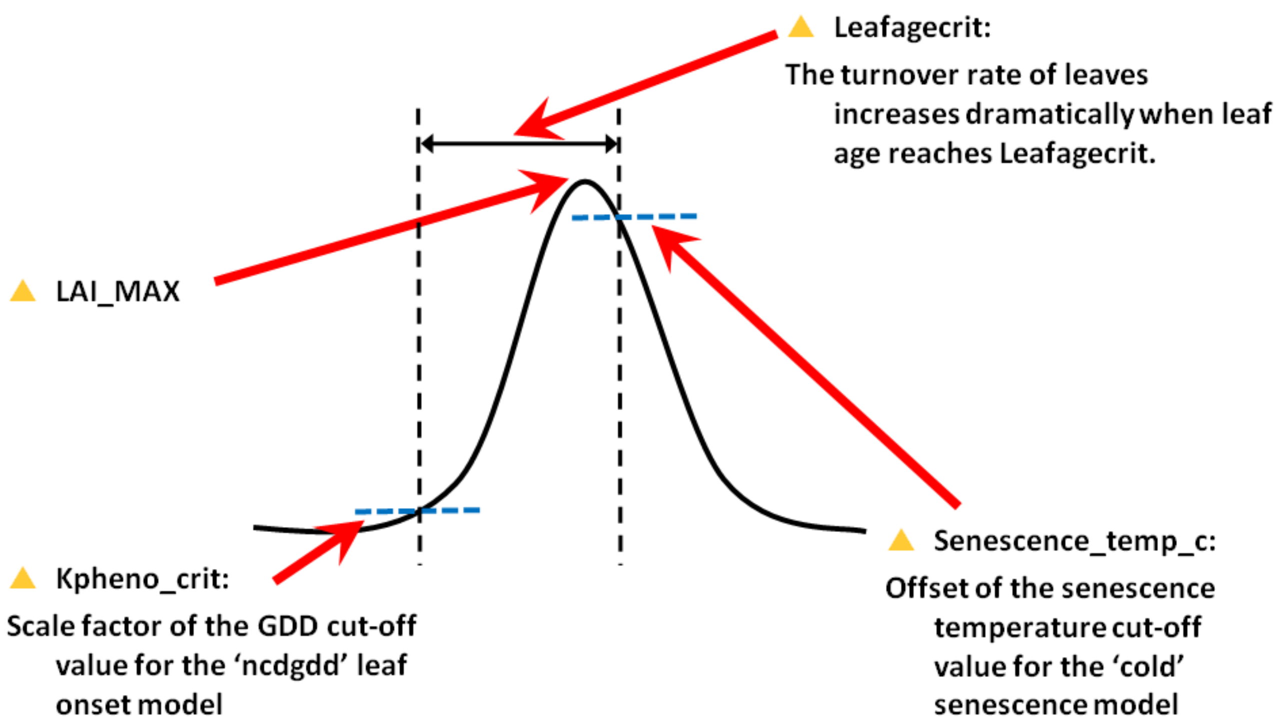
$$Rh = kresph \sum_s m_s B_s\, g(swc)(kQ10)^{\frac{Tsoil}{10}}$$

$$NEE = Rm + Rg + Rh - A$$

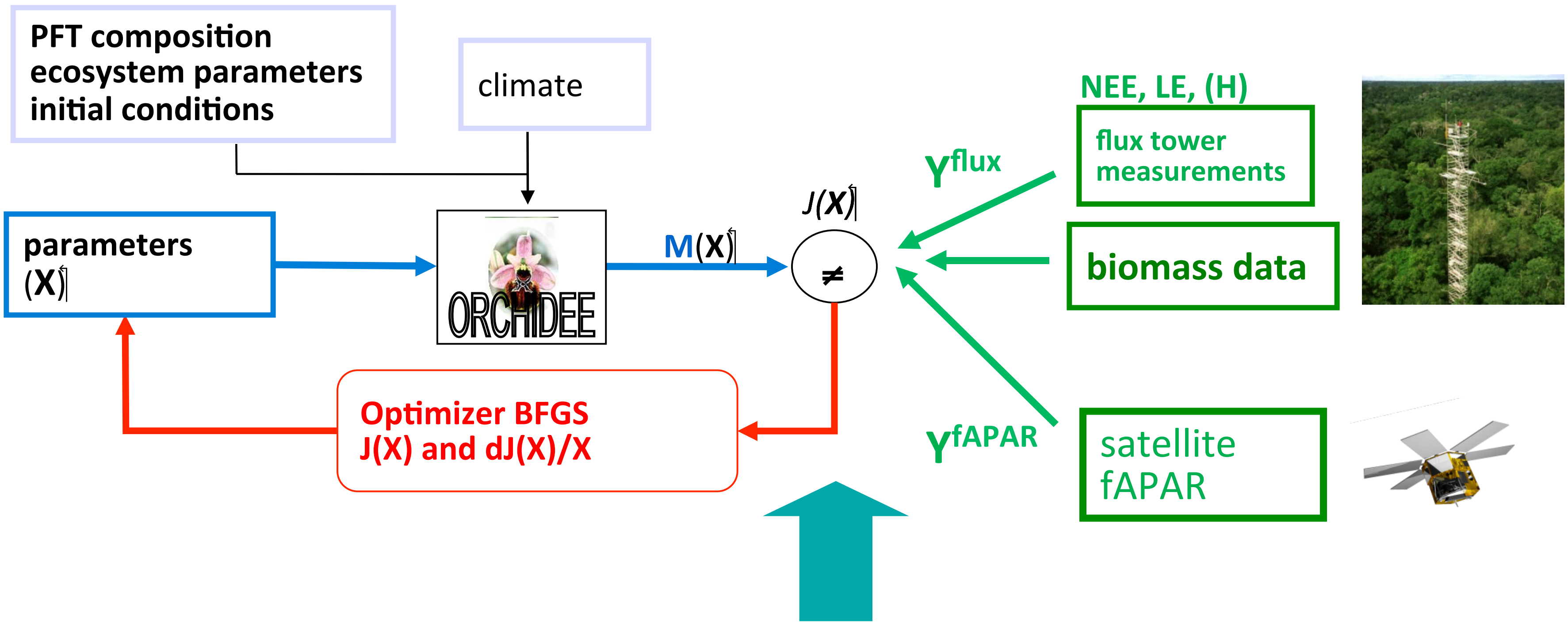
Biogeophysical equations are non linear with "many" thresholds

- **Example of the leaf phenology**

- **Tree Leaf onset based on a threshold (sum of growing degree days)**



Optimization implementation..

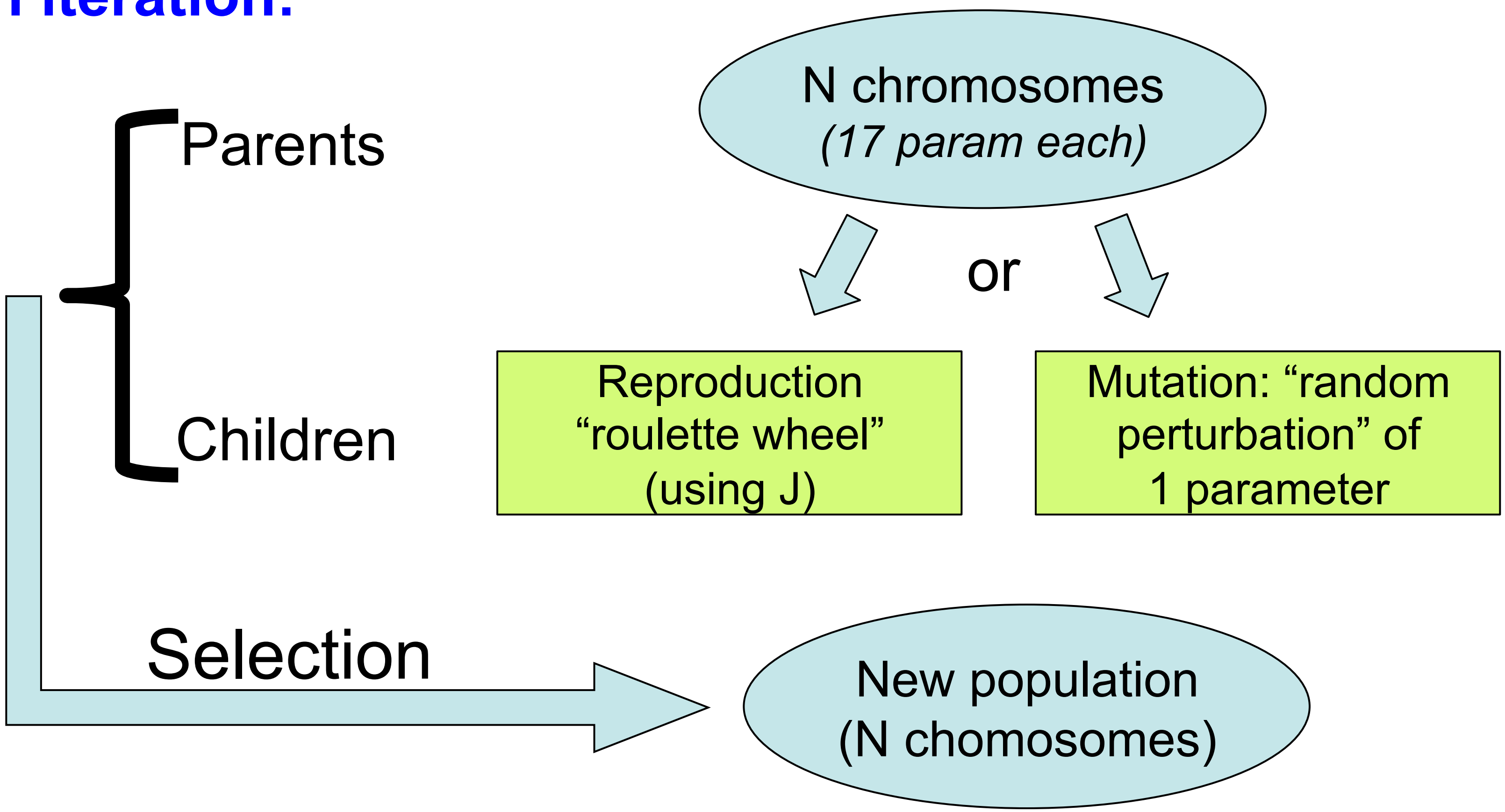


- **Cost function:** $J(x) = \frac{1}{2}\left[\left(y - M(x)\right)^t R^{-1}\left(y - M(x)\right) + (x - x_b)^t P_b^{-1}(x - x_b)\right]$

- **Iterative minimization using either:**
  - **Variational approach (with Tangent Linear model for DJ/dx)**
  - **Monte Carlo approach**

Genetic Algorithm set up..

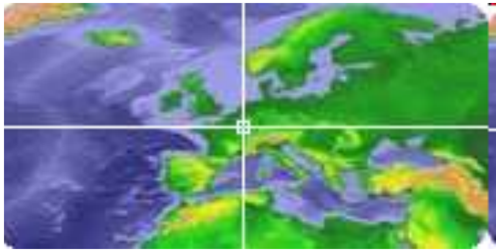
At each iteration:



- Mating rate = 80% ; Mutation rate = 20%
- Selection of population by elitism; 40 iterations

Assimilation of FluxNet data

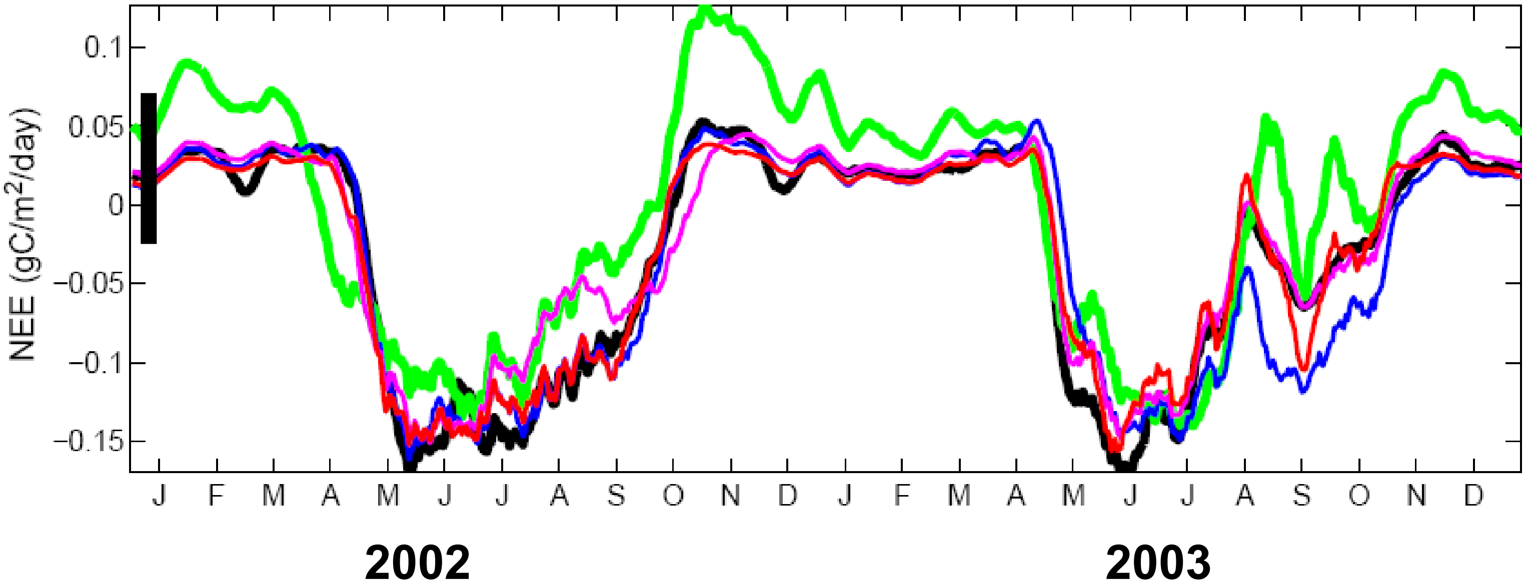
Question: ability of the model to represent year to year flux variations ?



- Ex: Beech forest : Hesse site (France)

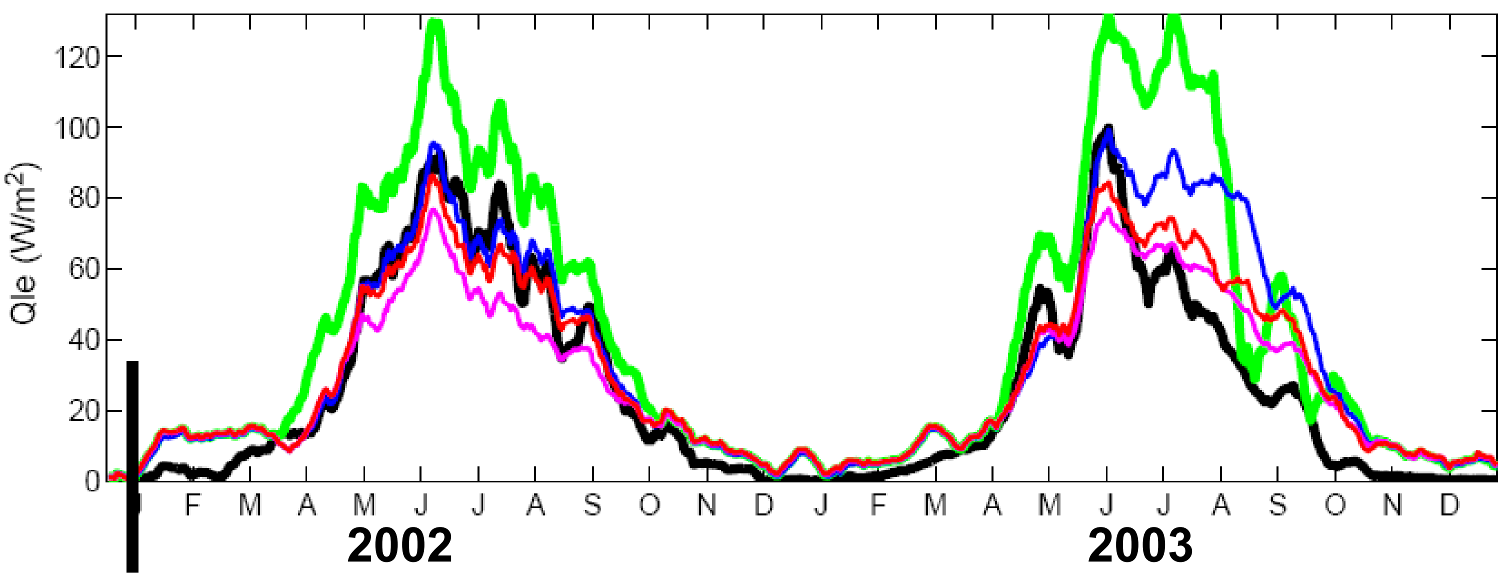
- Assimilation of daily NEE / Latent Heat
- 28 parameters
- period : 2001 – 2004
  - optimisation for each year separately
  - optimisation for the whole period

Hesse site : Model – data fit (2 yrs)



RMSE (x100)

Data
Prior: 41
2002: 22
2003: 19
All yrs: 13



RMSE

Data
Prior: 21
2002: 13
2003: 10
All yrs: 9

Ensemble vs Variational optimization

- Create Pseudo-Data with randomly perturbed parameters (within allowed range)

- Cost function (J) with no prior term

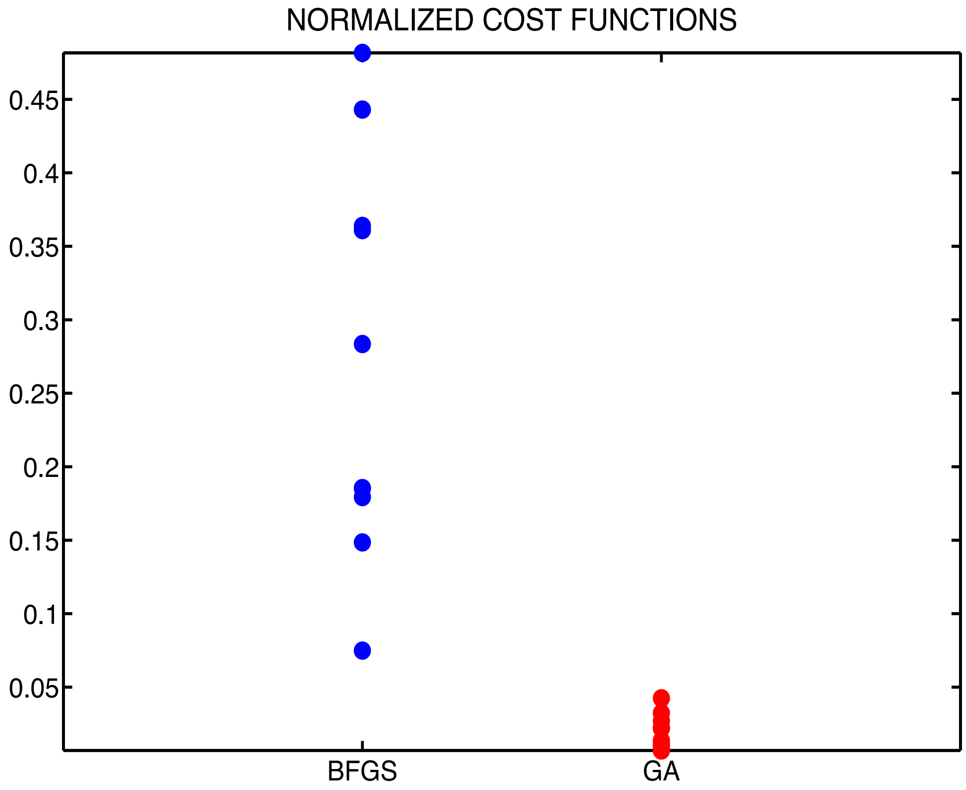
- 10 members
  - Variational scheme: 10 different first guest X
  - Genetic Algo. : 10 different experiments

- Use J_{ref} with ORCHIDEE standard param.

Genetic Algo. vs Variational : performance of the algorithm

- Random perturbation over 100% of the parameter range allowed variation

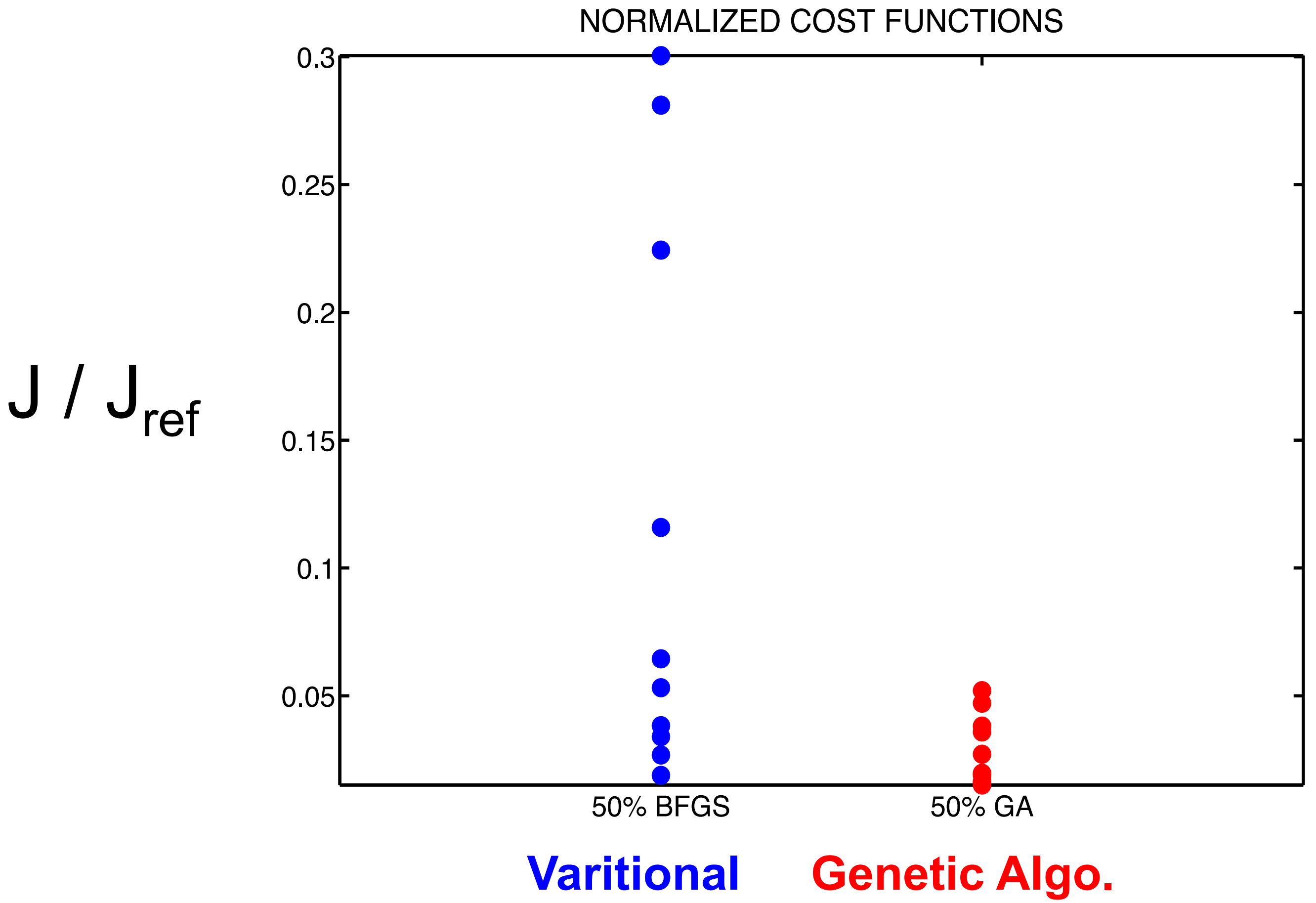
J / J_{ref}



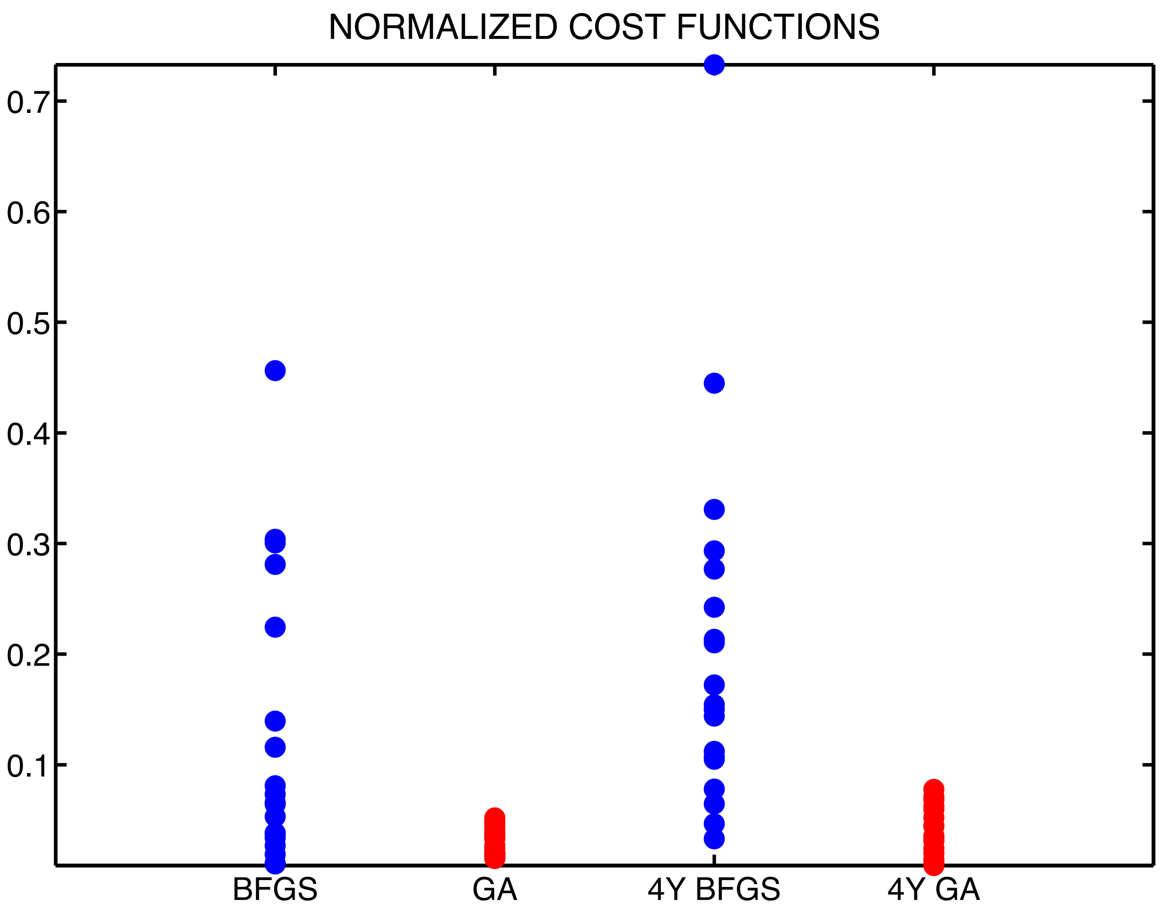
Varitional **Genetic Algo.**

Genetic Algo. vs Variational : performance of the algorithm

- Random perturbation over 50% of the parameter range allowed variation



Test with 4 yr vs 1 yr of data

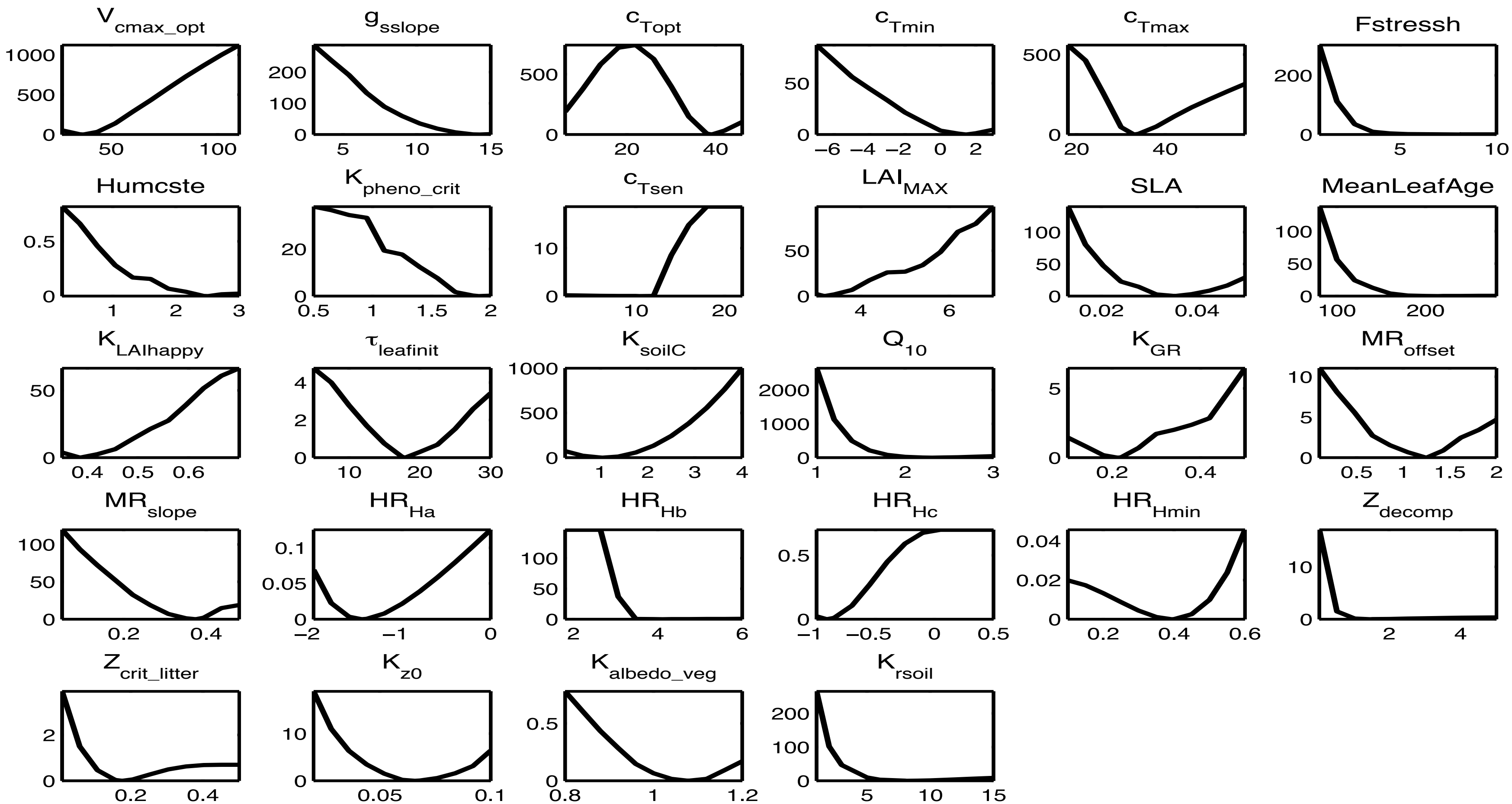


J / J_{ref}

1 year of data ***4 year of data***

Varitional **Genetic Algo.**

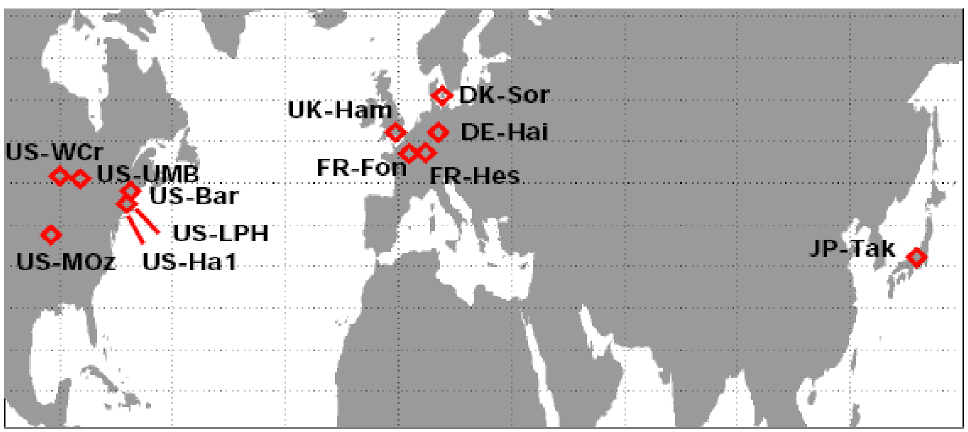
Cross section of J at "True" parameters (used for pseudo-data)



X range: Allowed range of variation for each parameter

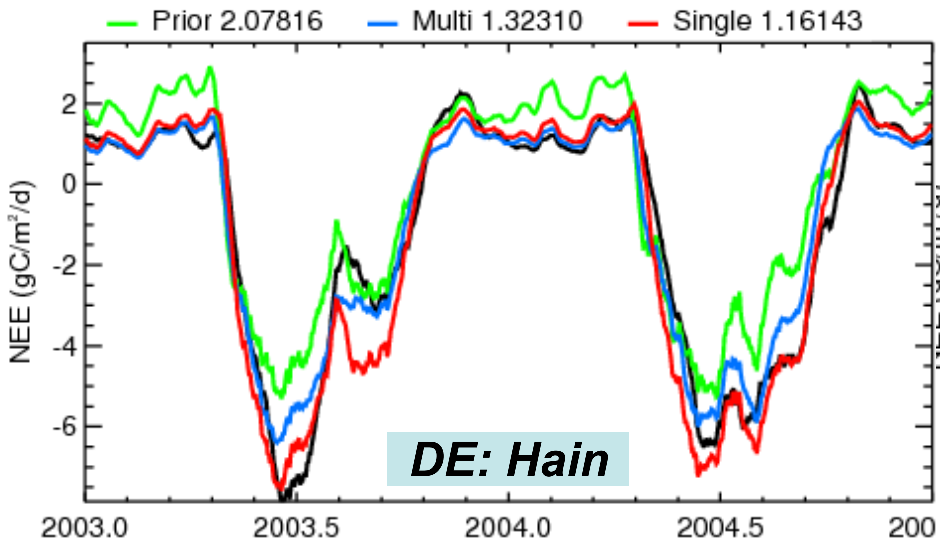
Comparative performances with more site-observations ?

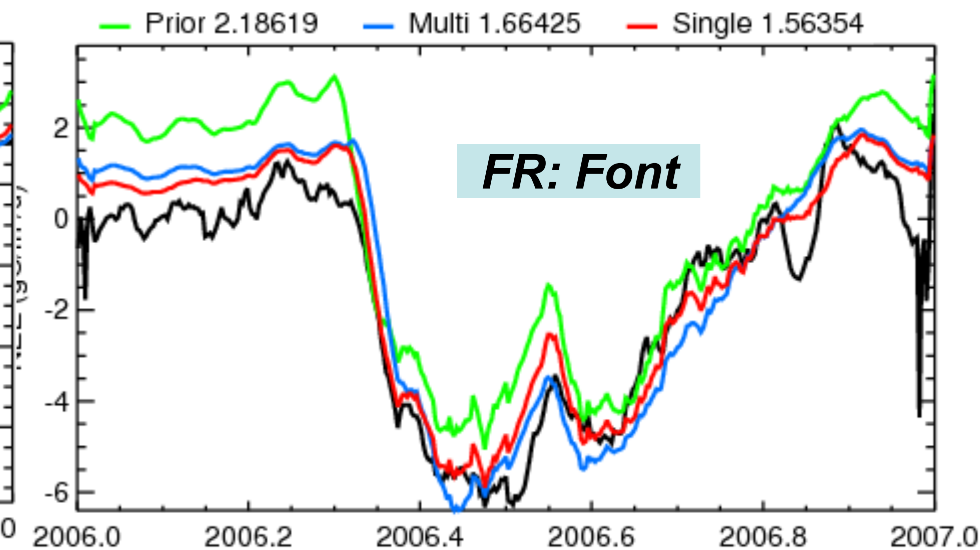
- How does the minimization performance change with more site-data in J ?
- 11 sites for a given Plant Functional Type
  Use observed daily measurements (NEE, LE)

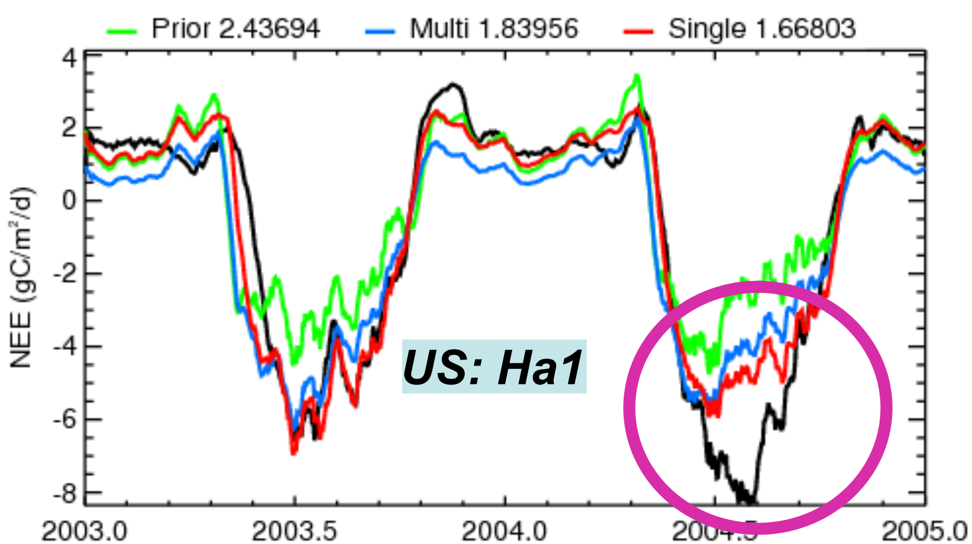


- First test only with Variational scheme
  (9 optimization with 9 random first guest)

Model – FluxNet data fit : ex. for 3 sites







NEE (gC/m2/d)

Data

Prior

Single-site posterior

Multi-site posterior

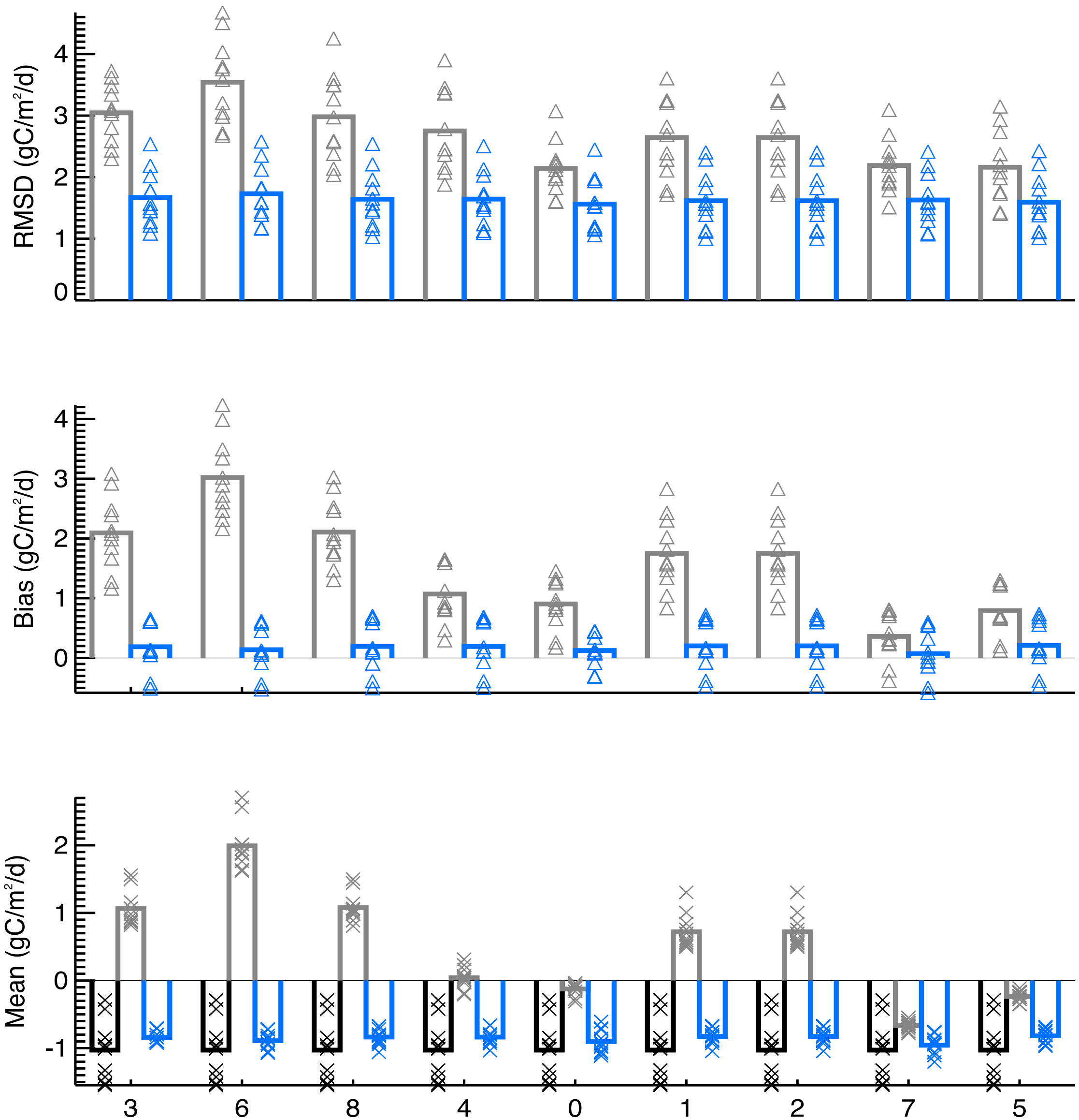
Kuppel et al. 2012

Results for 9 Variational optimizations

RMSE

Bias

Mean flux



Prior model

Posterior model

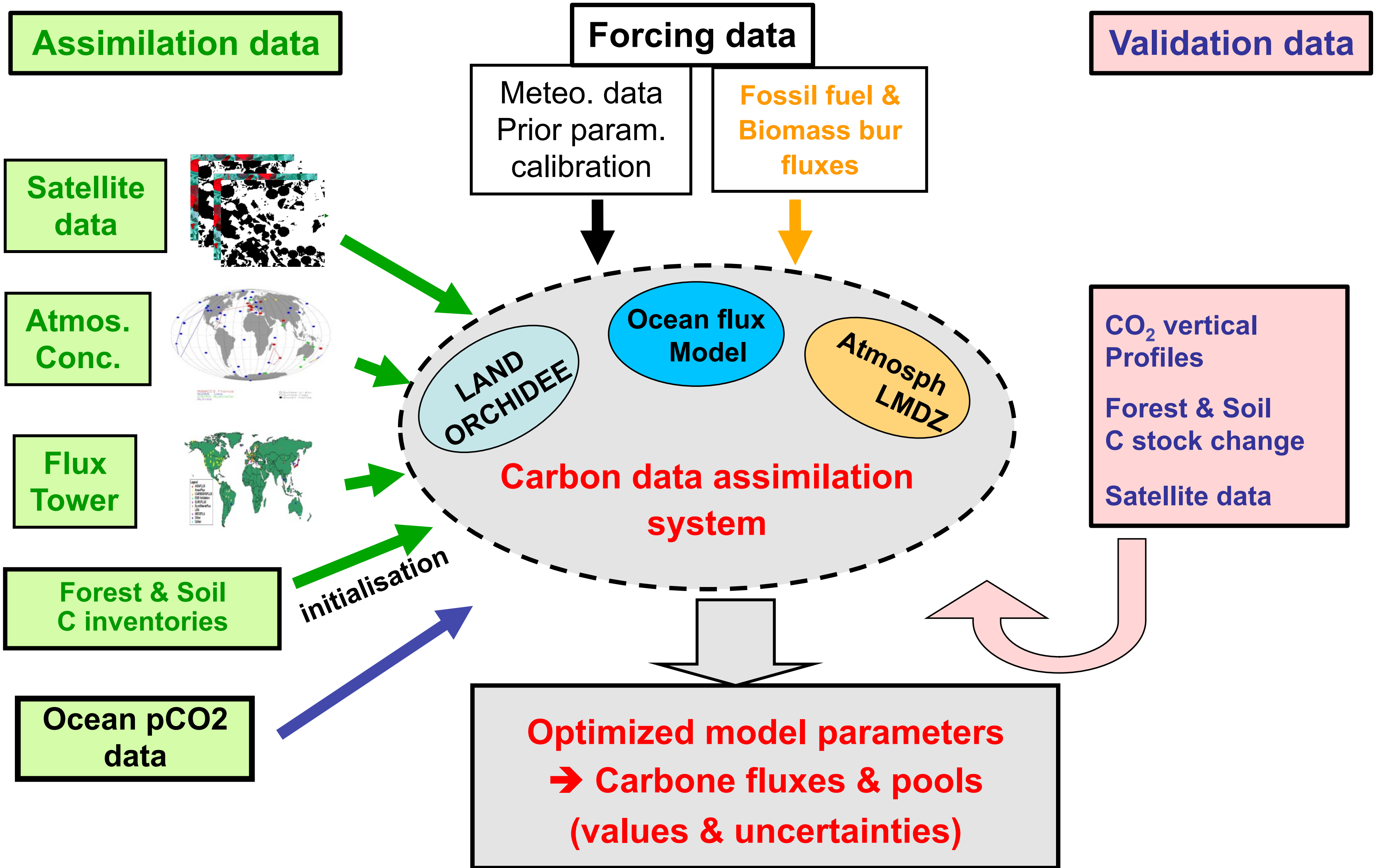
Observation

Conclusions…

➔ Data Assimilation is promising to improve simulations of the C cycle

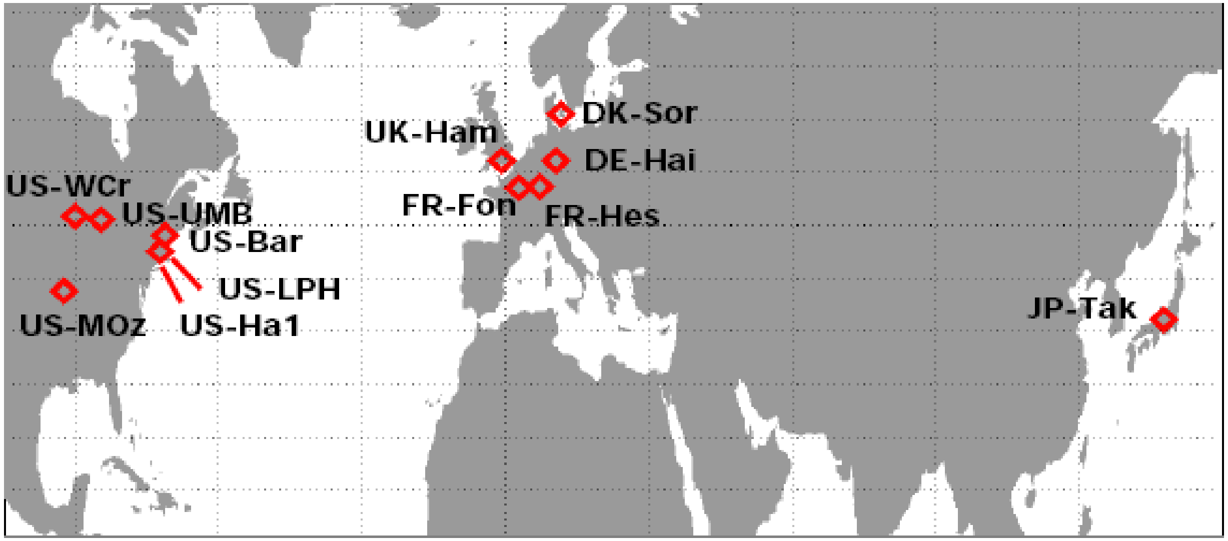
- Thresholds and non linearities complicate the optimization
  ➔ Need ensemble variational optim.
- Observation error matrix difficult to assess (i.e., model error)
- Need robust scheme easy to implement
- Variational methods based on Adjoint model are difficult to maintain…
  ➔ GA is promizing..
- Expert in DA are welcome in the C-cycle community…

Structure of a global "CCDAS"



Assimilation of Multi-site flux data

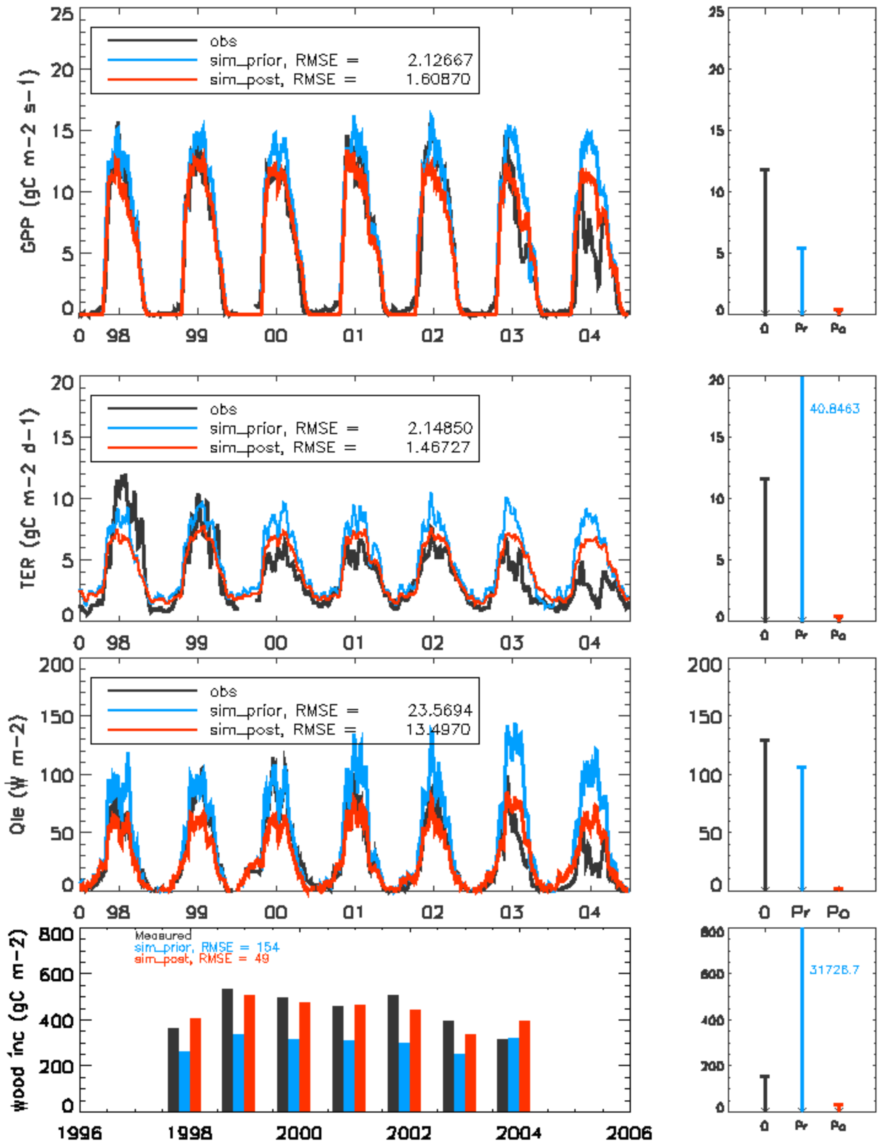
Ex: temperate Deciduous Broadleaf Forest
use 12 sites with > 70 % DBF coverage



- **Obs type :** **NEE & Latent heat**
- **Resolution :** **daily data**
- **period :** **3 to 4 years per site**

Kuppel et al. 2012

Assimilation of Biomass measurements (ex: Site level ; Beech Forest ; France)



FluxNet data (daily) &Yearly Biomass increments

(25 flux related params)

Measurement

Prior model

$GPP_{RMSE} = 2.1$

$TER_{RMSE} = 2.1$

$Qle_{RMSE} = 24$

Posterior model

$GPP_{RMSE} = 1.6$

$TER_{RMSE} = 1.4$

$Qle_{RMSE} = 13$

Thum et al. 2012

Summary: Potential of a CCDAS..

➔ Promizing approach to account for multi-data streams

➔ Helps to identify model deficiencies !

➔ Relative Error characterization bw data stream becomes crucial for a proper assimilation

➔ Anticipated data streams to become crucial:
- soil-C observations…
- data from Ecosystem Manipulative Experiments

➔ Ongoing large community effort :
- GeoCarbon EU-project (5 land CCDAS)
- Existing inter-comparison of Model-Data fusion exercise

Limitations of a CCDAS…

➔ Strongly rely on a given model structure

➔ Missing processes in the ecosystem model might lead to
- Wrong parameter estimates
- Poor model predictability (Strong biases)

➔ Non-linearities might complicate the parameter optimization

➔ Need to :
- keep independent data for model output validation
- Keep classical Atmospheric inversion

Formalism…

$$J(\mathbf{x}) = (\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + (\mathbf{Hx}-\mathbf{y})^T\mathbf{R}^{-1}(\mathbf{Hx}-\mathbf{y})$$

- **Analytical solution**
  - Need to linearize the model H(x)
  - Sensitivities (H) from tangent linear or Adjoint

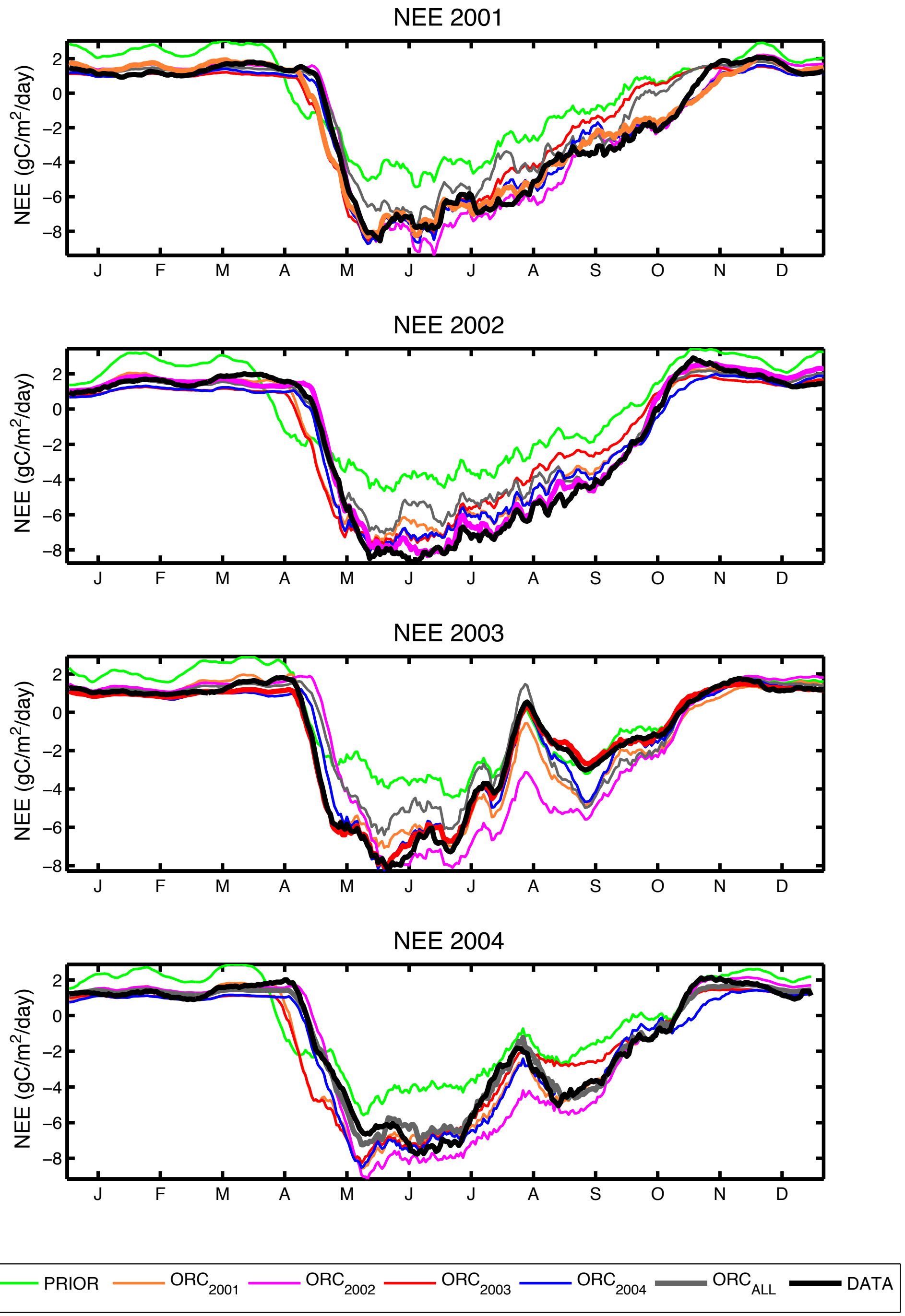
$K = (H^T.R^{-1}.H + B^{-1})^{-1}\ H^T.R^{-1}$

$$\mathbf{K} = \mathbf{BH}^T(\mathbf{HBH}^T+\mathbf{R})^{-1}$$
$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y}-\mathbf{Hx}_b)$$
$$\mathbf{A} = \mathbf{B} - \mathbf{KHB}$$

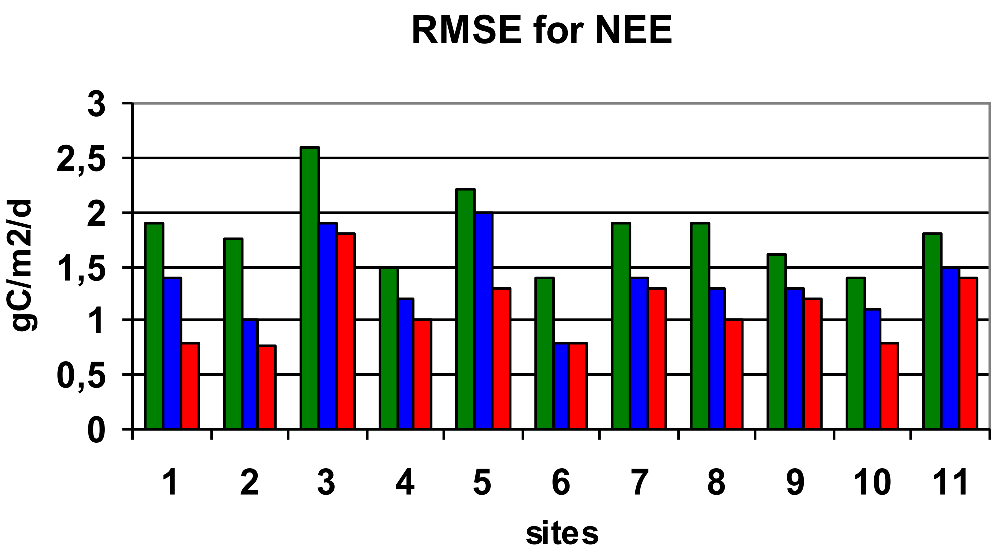
- **Variational solution**
  - Adapted to large size problems
  - Error estimation more difficult

$$\nabla J(\mathbf{x}) = 2\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + 2\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{Hx}-\mathbf{y})$$
$$\mathbf{A} = 2[J''(\mathbf{x}_a)]^{-1}$$

- **Monte Carlo approaches**
  - Used mostly for site-level studies
  - Required time may be prohibitive with “complex model”
  - No limitations wrt LINEARITY & parameter PDF


NEE 2001
NEE 2002
NEE 2003
NEE 2004
NEE (gC/m^2/day)
2
0
−2
−4
−6
−8
J F M A M J J A S O N D
PRIOR
ORC$_{2001}$
ORC$_{2002}$
ORC$_{2003}$
ORC$_{2004}$
ORC$_{ALL}$
DATA

Model-data fit for NEE (RMSE)



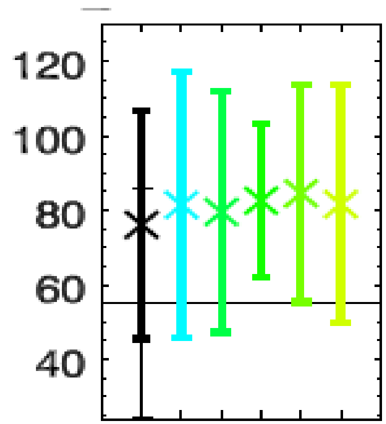
Prior   Single-site posterior
Multi-site posterior

➔ RMSE reduce by
- 30% for NEE
- 15 % for LE

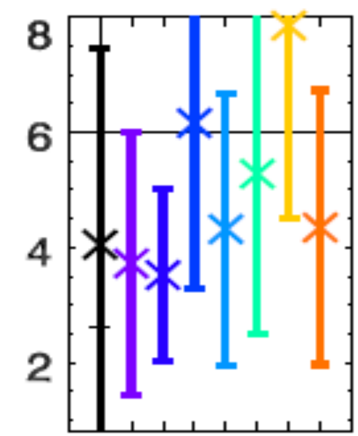
Kuppel et al. 2012

Retrieved parameter values

Maximum photosynthetic capacity



Water stress function (slope)



Black: Multi-site

Colors: Single-site

➔Agreement betw all sites

➔Need to improve model !

Kuppel et al. 2012