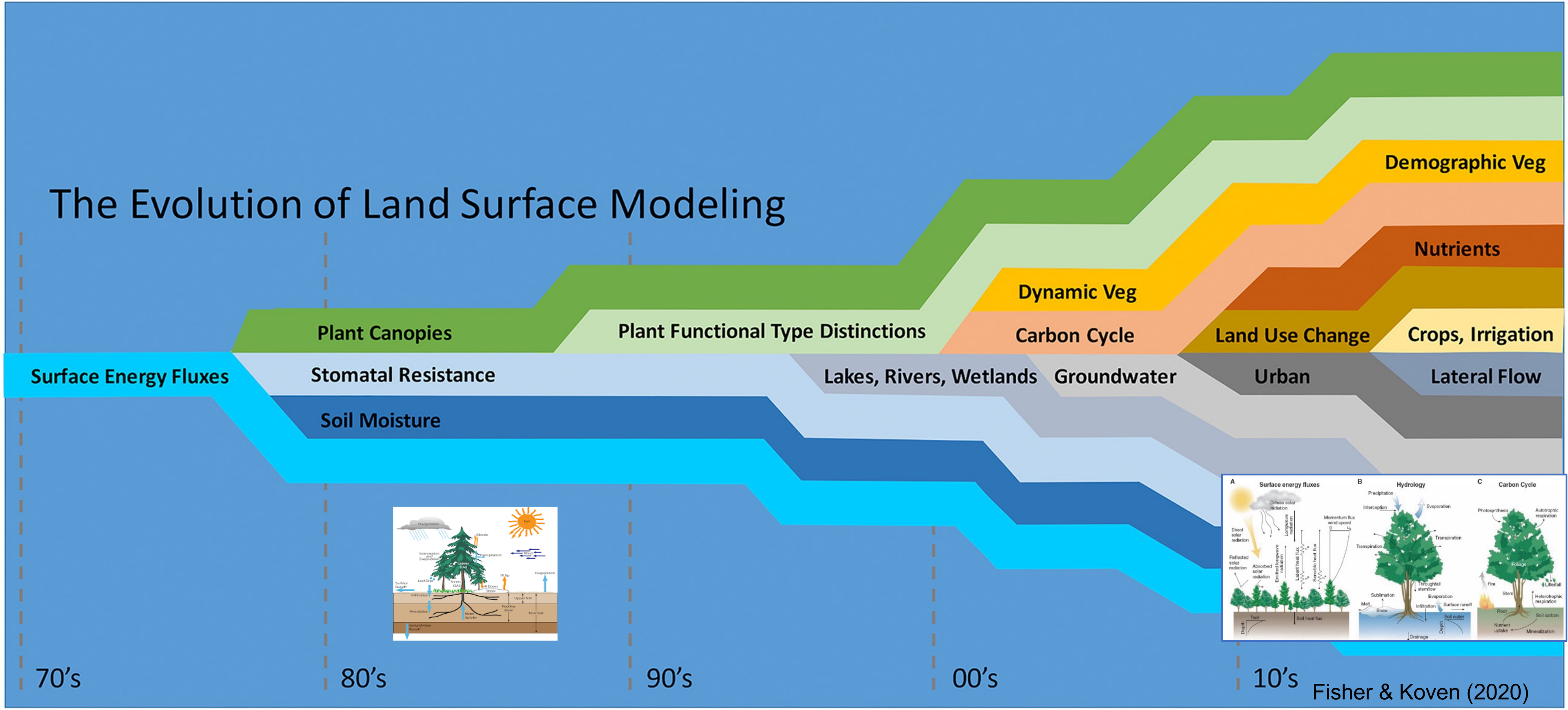
Global land surface model parameter optimisation: where we are and new opportunities ? Examples from the ORCHIDEE model

Philippe Peylin *with direct contributions from*
Natasha McBean, Cedric Bacour, Vladislav Bastrikov, Nina Raoult, Catherine Ottle, Fabienne Maignan, Simon Beylat, Sylvain Kuppel, Nuno Carvalhais, ...

Outline...

- Motivation for Data Assimilation (parameter calibration)
- Highlights of scientific results and issues linked to parameter optimisation
- Remaining key challenges & Upcoming opportunities

History of global land surface modeling !

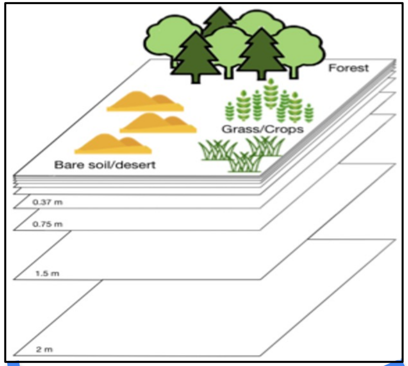


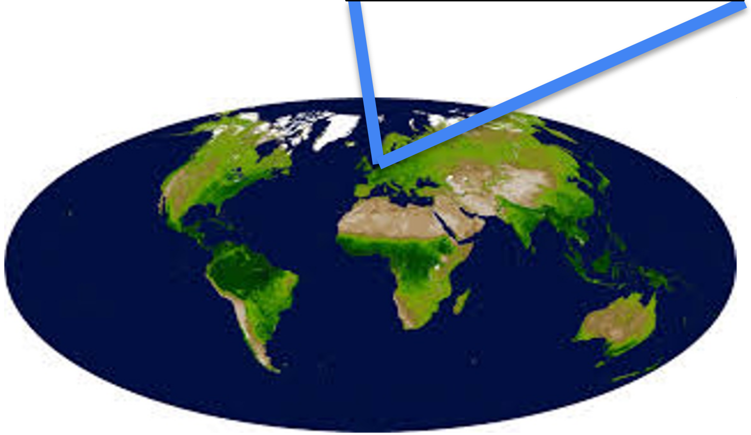
Fisher & Koven (2020)

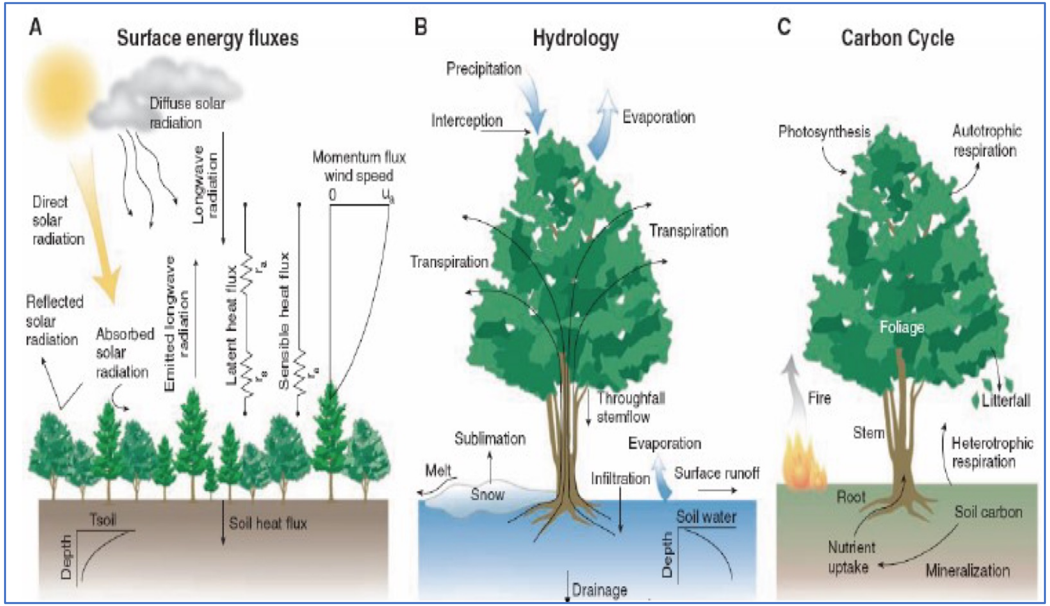
Land surface model

⇒ Solve for Energy / Water / Carbon / Nitrogen budgets

Global land surface models

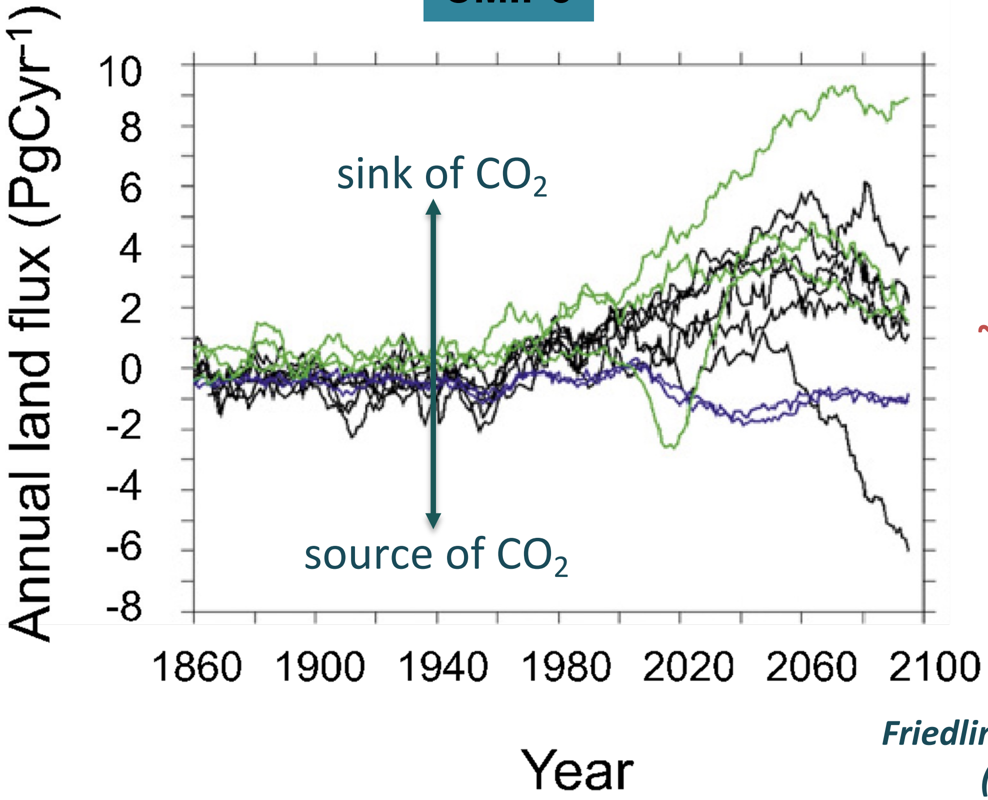






URGENT need to reduce uncertainty in global carbon sink projections!

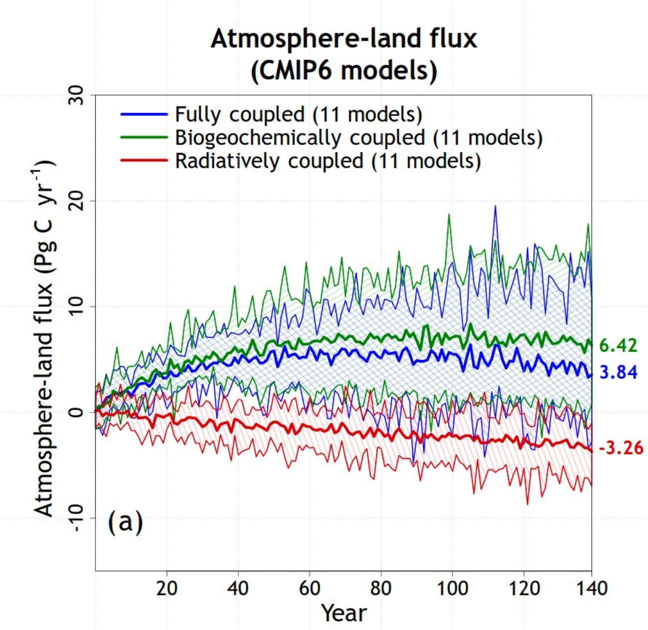
CMIP5



~14.5PgC

Friedlingstein et al. (2014)

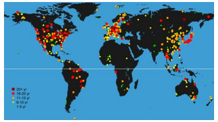
CMIP6



Vivek et al. (2020)

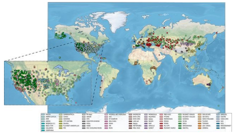
Recent large increase of available observations !

Large / Numerous *in situ* data networks

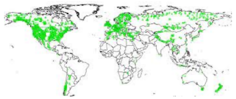


FluxNet measurements
Soil Chamber measurements

Manipulative experiments
(ex. FACE, …)



Surface Soil Moisture
Network



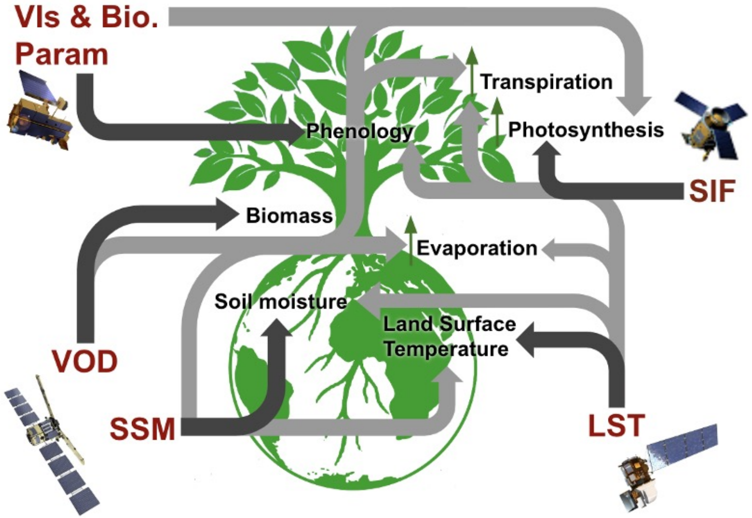
International Tree Ring
database (ITRDB)

TERENO
TERRESTRIAL ENVIRONMENTAL OBSERVATORIES
Data Discovery Portal

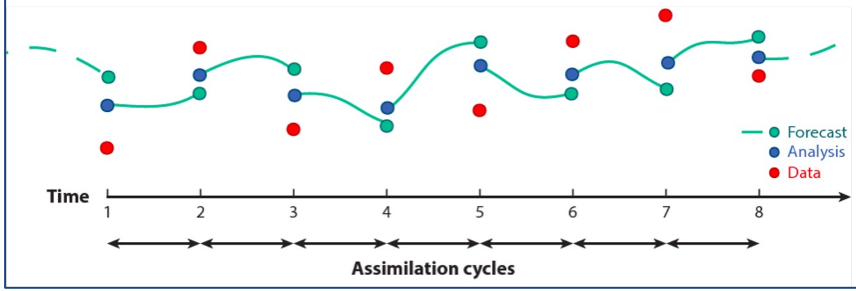
OZCAR
eLTER FRANCE

Satellite observations

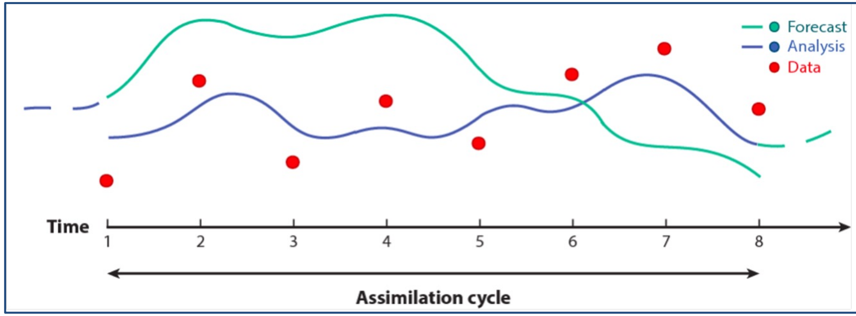
Increasing data stream
large increase in spatial resolution



We can use **data assimilation** to **reduce parameter uncertainty** in global land surface models



Sequential:
one at a time
common in state estimation



Variational:
full window
common in parameter estimation

Bayesian cost function

Vector of parameters

Mismatch between the observations and the model

“Background” parameters i.e. default parameter values

$$J(\boldsymbol{x}) = \tfrac{1}{2}(\ \boldsymbol{y}-M(\boldsymbol{x})\ )^{T}\mathbf{R}^{-1}(\ \boldsymbol{y}-M(\boldsymbol{x})\ ) + \tfrac{1}{2}(\ \boldsymbol{x}-\boldsymbol{x}_{b})^{T}\mathbf{B}^{-1}(\ \boldsymbol{x}-\boldsymbol{x}_{b})$$

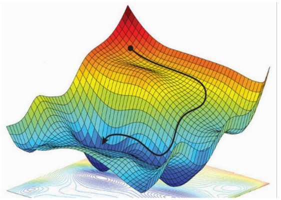
Observations
e.g. data from International Soil Moisture Network

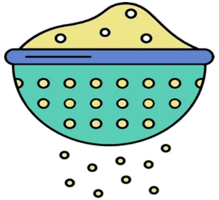
Model output given the set of parameters **x**
e.g. modelled soil moisture

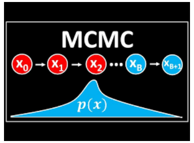
Error covariance matrix

Mismatch between the parameters tested and their values

Various methods…







Gradient descent

Particle filters

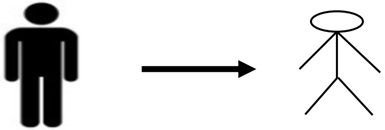
Genetic algorithm

MCMC

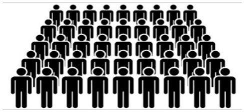
……..

Various algorithm…

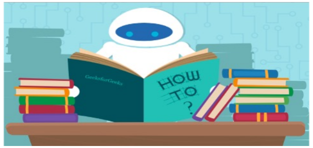
- Tangent linear / Adjoint $\frac{dy}{dx} = f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$
- Emulation



- Ensembles



- Machine learning



Outline…

- Motivation for Data Assimilation (parameter calibration)
- **Highlights of scientific results and issues linked to parameter optimisation (biased with examples from the ORCHIDEE model)**
- Remaining key challenges & Upcoming opportunities

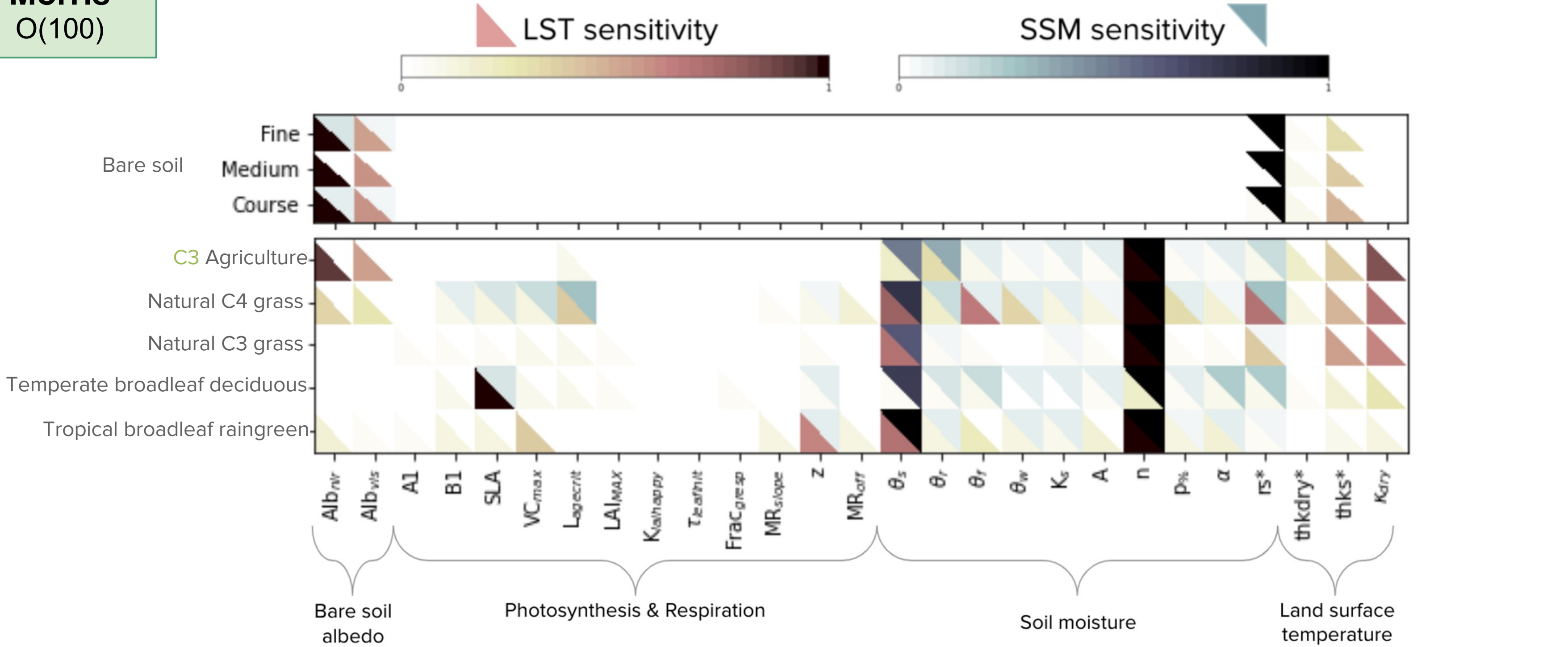
Which parameters to optimise ?

Need to use parameter sensitivity analysis

➔ Morris' method allows us to identify to the most sensitive parameters !

Morris
O(100)

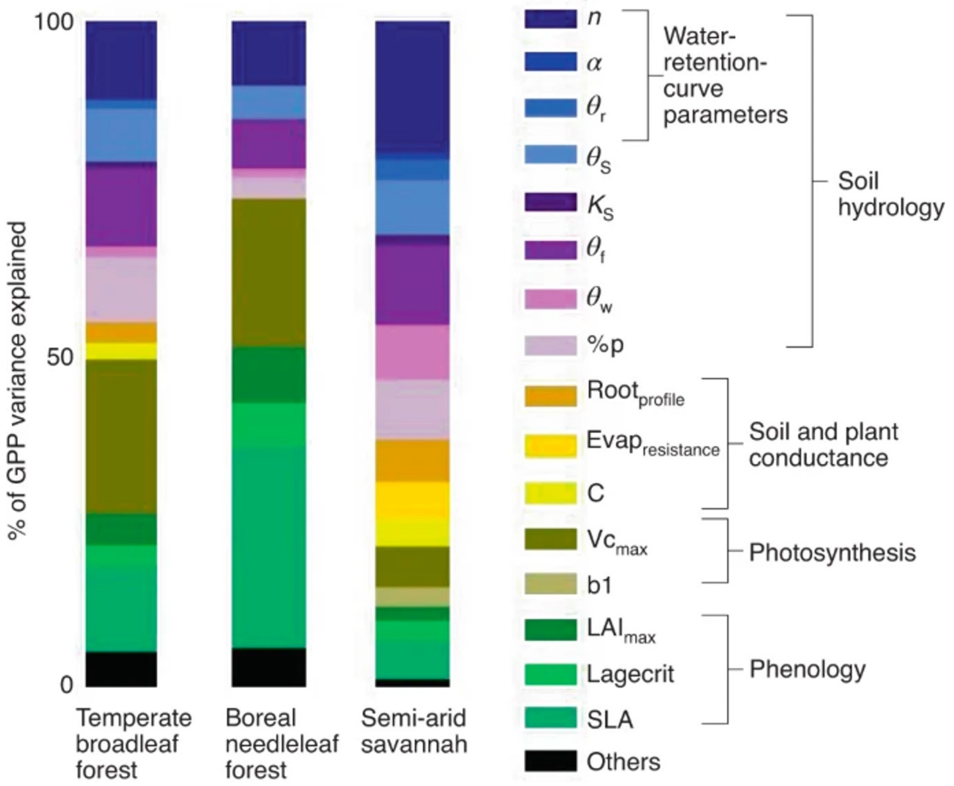
ORCHIDEE Parameters for Surface Temperature / Surface soil Moisture



Need to use parameter sensitivity analysis

Sobol's method allows to capture the **interactions** between the parameters

Sobol
O(10,000)



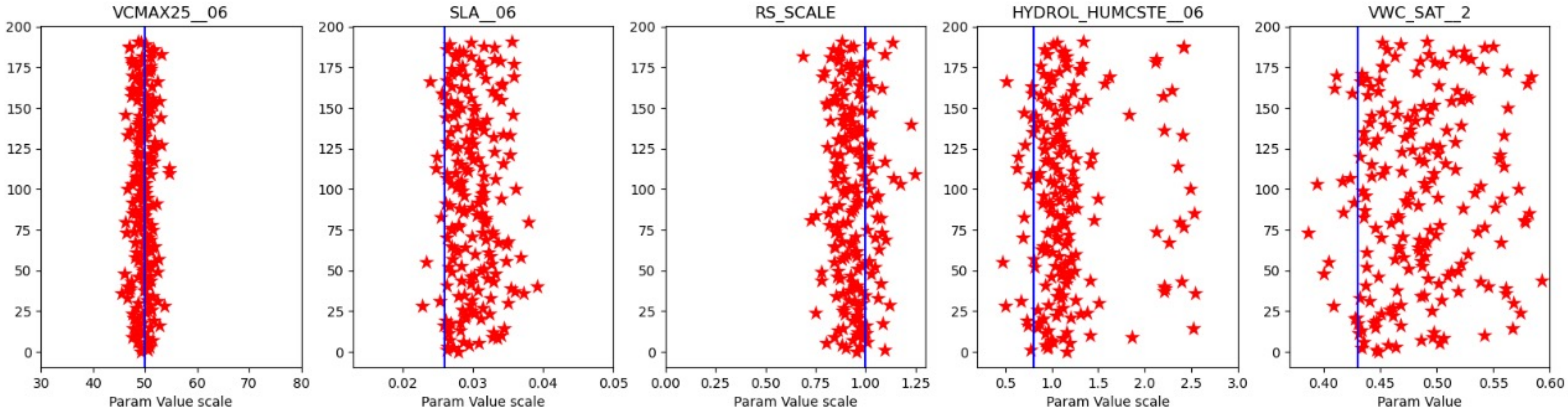
Novick et al. (2022)

Pseudo-data experiment (with known True parameters) !

➔ Pseudo-data experiments is highly recommended !

- Create peusdo – obs with perturbed parameters
- Try to retrieve the True param starting with different first Guess !

Example: ORCHIDEE model, 5 parameters ; 1 year of GPP pseudo data ; 200 first-guess tests



True Param Posterior Param

Which metrics (for a given data stream) ?

Which metrics and which cost function ?

Observation operator:

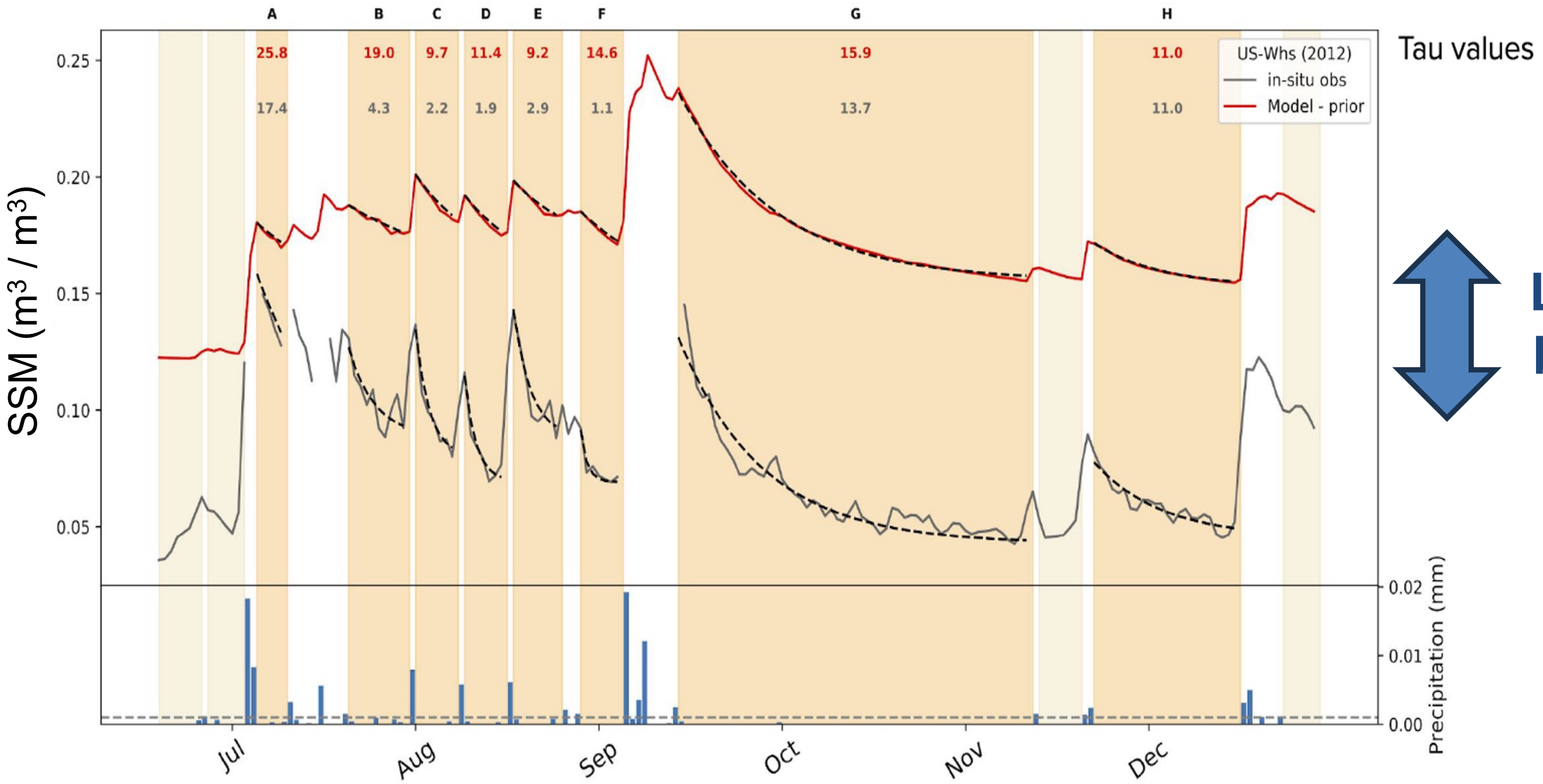
- Need for robust operator (spatially & temporally)
- Need to minimize the influence of model and observation **biases** !
- Need to characterize accurately **model and observation errors** as well as error correlations

Cost function :

- Quadratic or least absolute value or ??

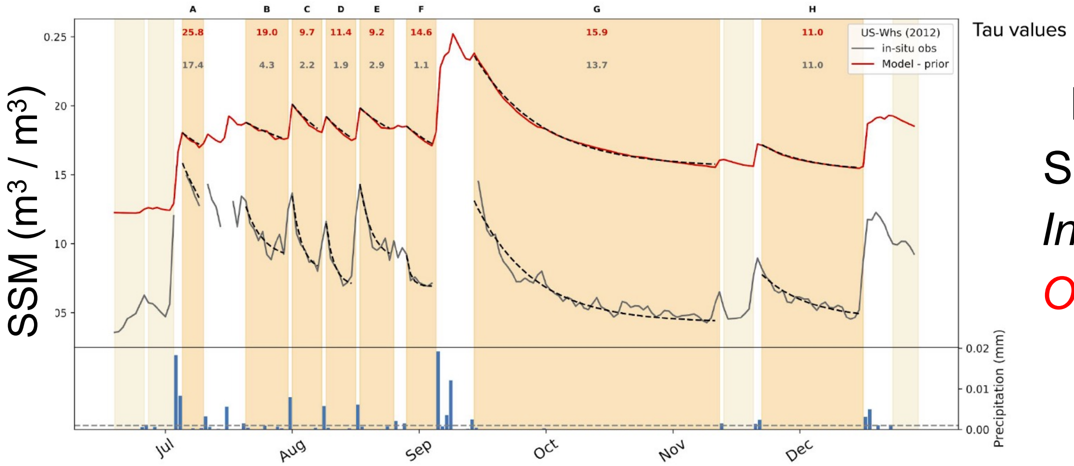
Case study with Surface Soil Moisture

Example (1 site): **In situ Surf. Soil Moisture data** / ***ORCHIDEE** model SSM*



Large biais !

Case study with Surface Soil Moisture

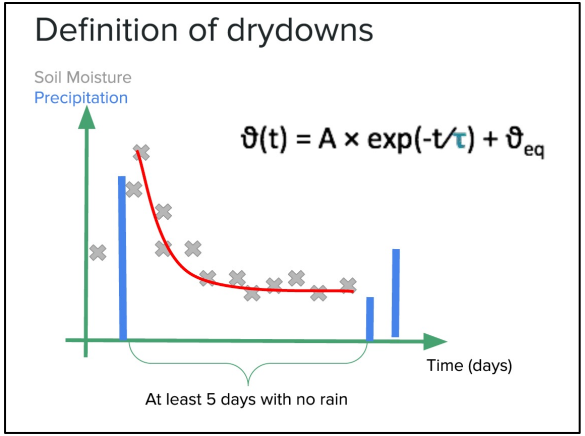


Example (1 site)

Surf. Soil Moisture:

In situ SSM

ORCHIDEE SSM



Treat biases using either

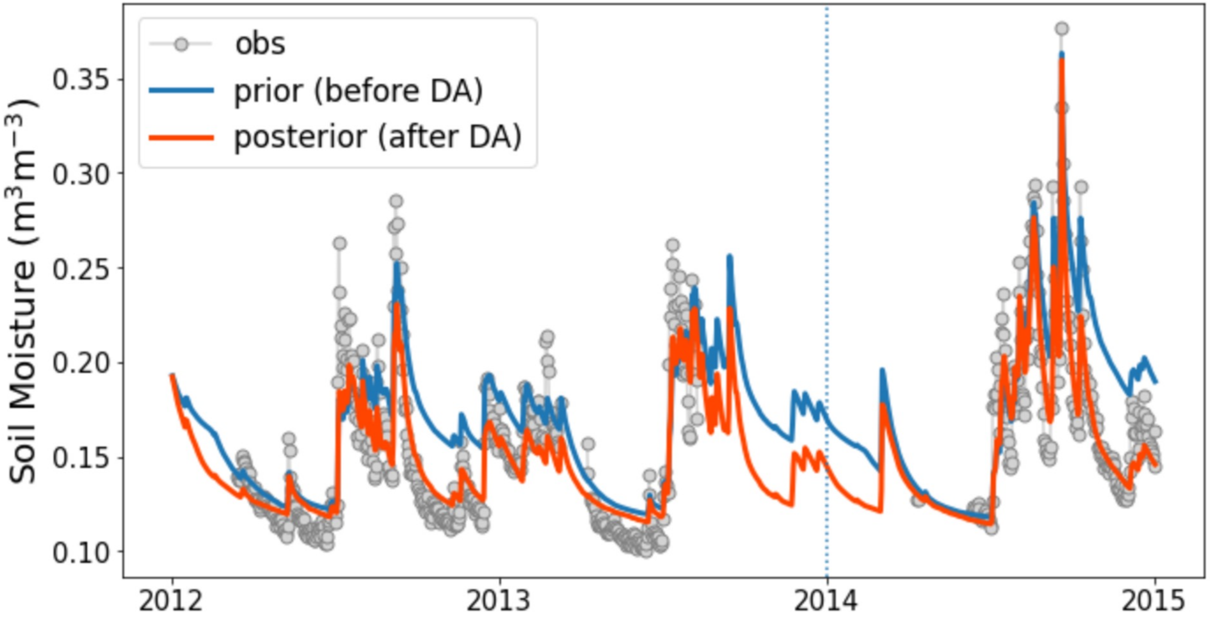
- ➢ Unbiased RMSE
- ➢ or CDF matching of SSM data
- ➢ or measure of dry down rate (exponential fit "Tau")

➔ Depend on the objectives

Optimisation

Calibration:
32% RMSD reduction

Evaluation:
49% RMSD reduction



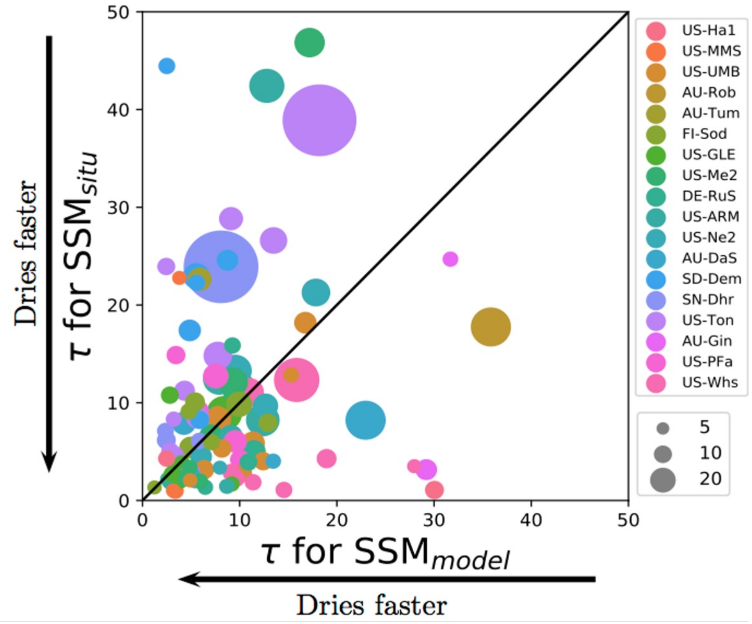
Example over US-Whs: Walnut Gulch Lucky Hills Shrub

- Optimised **τ** values
- Bias corrected runs shown
- 37% improvement in correlation over the whole period

Raoult et al. (2020)

Optimization with in situ data

⇒ Calibration against “Tau” (or “raw SSM”) at ‘18 sites with SSM & FluxNet data’



Length of dry period

- Model tends to dry out faster/slower depending on sites !
- Too small sample of sites to conclude about vegetation, soil texture or climate factors

Raoult et al. (2021)

Single vs Multiple sites optimisation ?

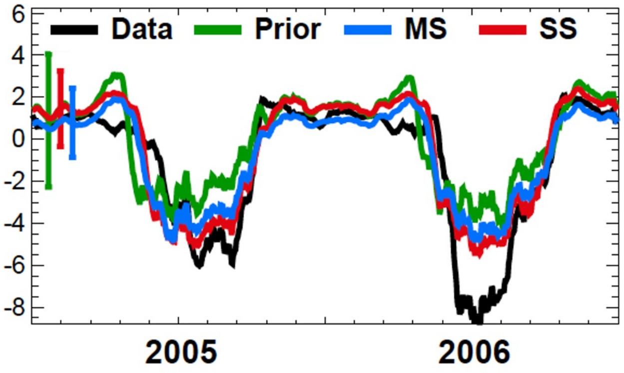
Single site vs Multi site optimization

Ex: Harward forest : Temperate Broadleaf decidious forest (12 sites)

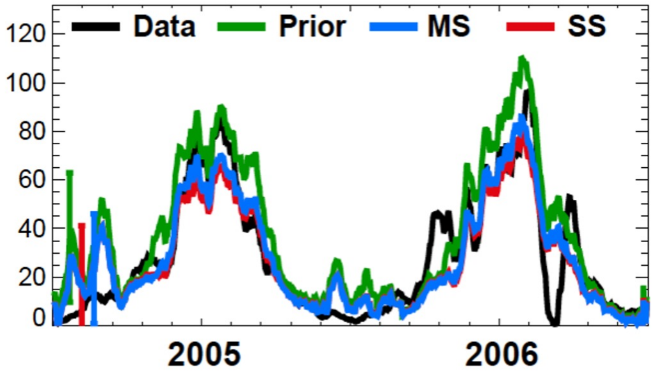
Assimilation
Of NEE and LE
fluxnet data

≈ 15 params / PFT

NEE (gC/m2/d)



LE (w/m2)



Data Prior Single site optim Multi site optim (15 sites)

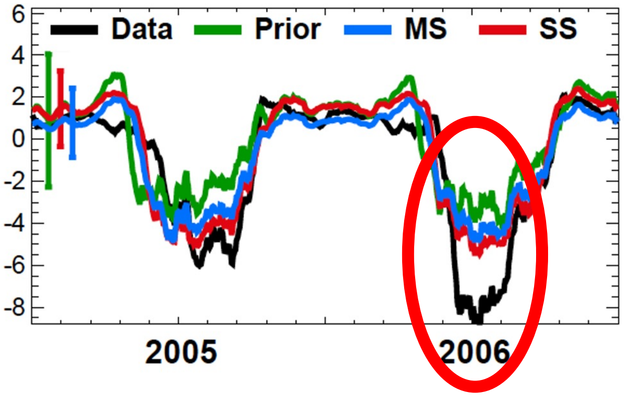
Single site vs Multi site optimization

Ex: Harward forest : Temperate Broadleaf decidious forest (12 sites)

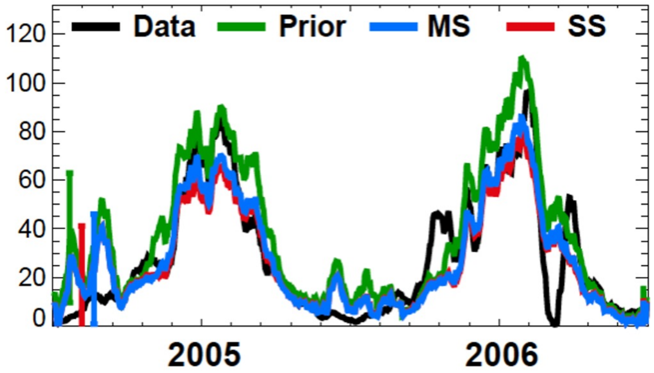
Assimilation
Of NEE and LE
fluxnet data

≈ 15 params / PFT

NEE (gC/m2/d)



LE (w/m2)



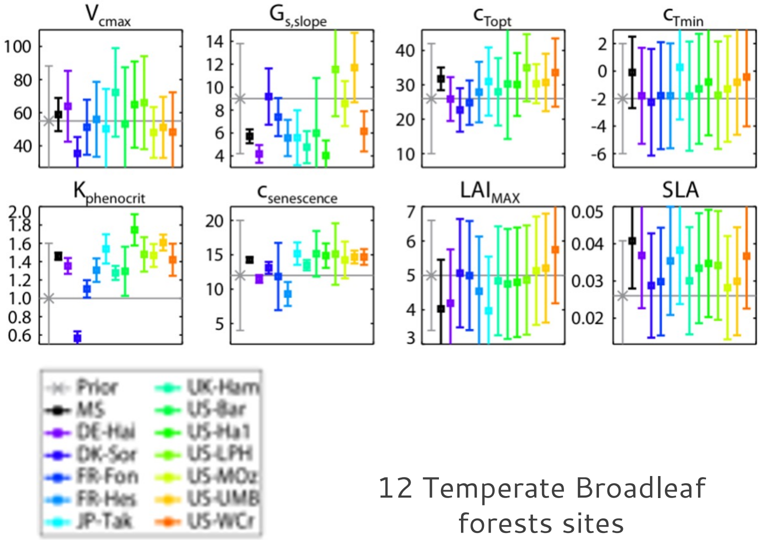
Data Prior Single site optim Multi site optim (15 sites)

➔ DA to highlight model deficiencies !

Single site vs Multi sites optimisation

Variability of the parameter estimates with site

➔ **Multi-site parameter values (black symbols) are often NOT the mean of the single site values (colored symbols) !**



12 Temperate Broadleaf forests sites

@Kuppel et al. (2012)

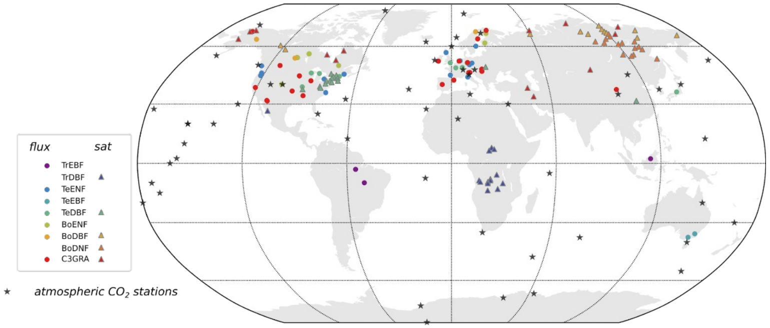
Single vs Multiple data-stream assimilation

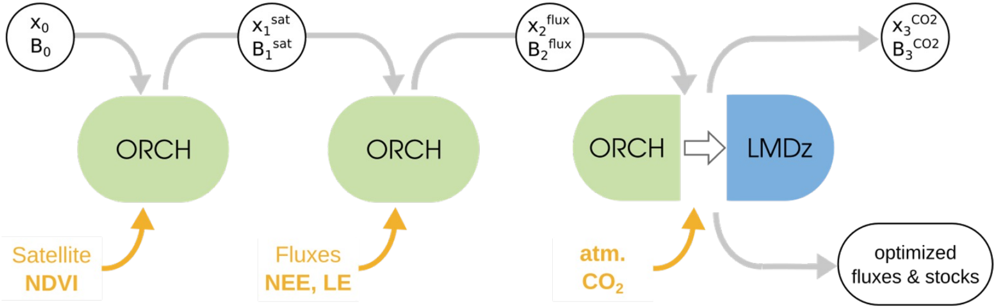
Multiple constraints on global carbon stocks and fluxes

Satellite NDVI

FluxNet C/W

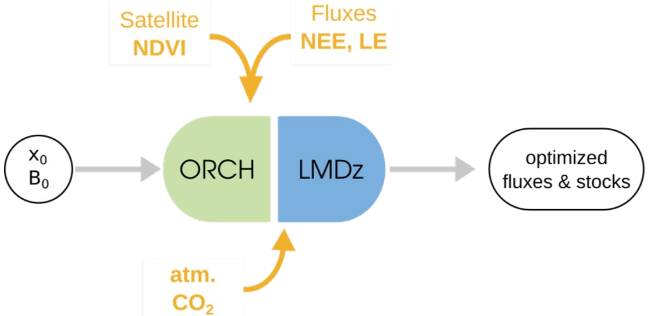
Atm. CO2





Sequential approach

@Peylin et al. (2016)



Simultaneous approach

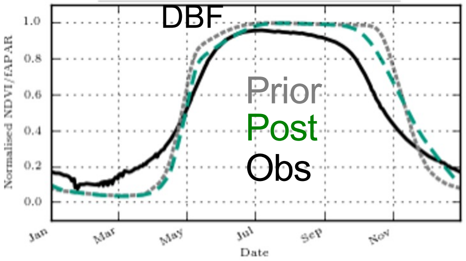
@Bacour et al. (2023)

Multiple data streams assimilation

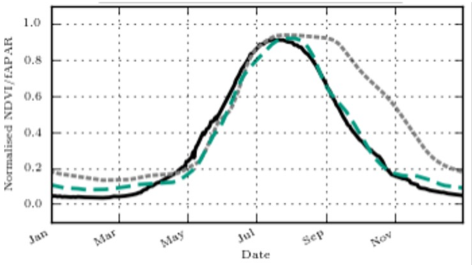
Step 1: MODIS-NDVI

4 params / PFT

Temp DBF



Bor DBF



➔ **NDVI is now often replaced by Solar Induced Fluorescence (OCO2; TROPOMI, GOME)**

Zeng et al. 2023

nature.com/articles/s41559-023-02187-6

nature ecology & evolution

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > nature ecology & evolution > articles > article

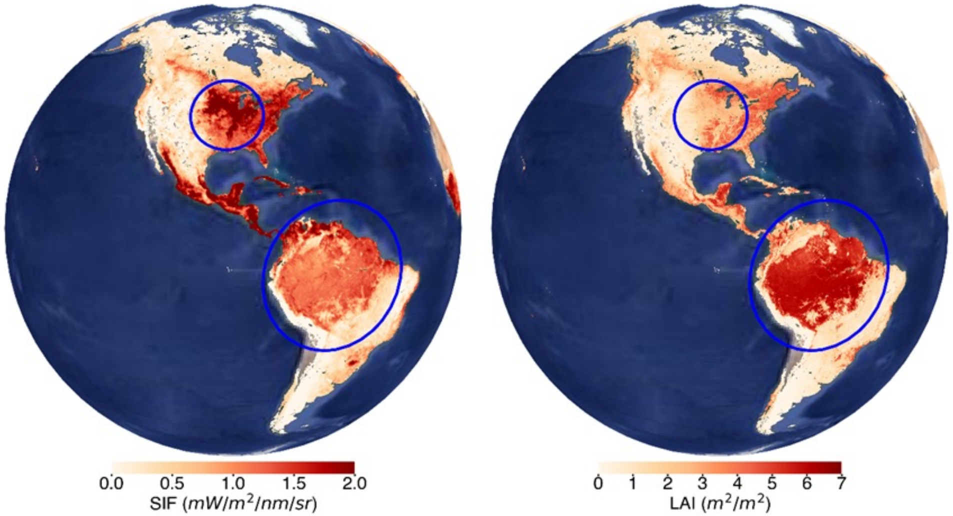
Article | Published: 14 September 2023

Structural complexity biases vegetation greenness measures

Yelu Zeng, Dalei Hao, Taejin Park, Peng Zhu, Alfredo Huete, Ranga Myneni, Yuri Knyazikhin, Jianbo Qi, Ramakrishna R. Nemani, Fa Li, Jianxi Huang, Yongyuan Gao, Baoguo Li, Fujiang Ji, Philipp Köhler, Christian Frankenberg, Joseph A. Berry & Min Chen

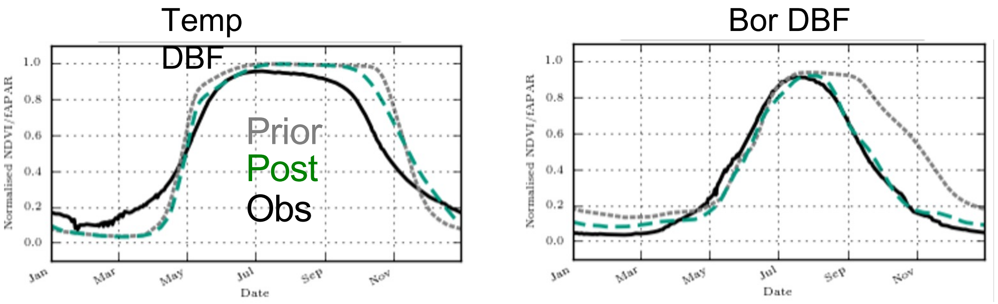
Nature Ecology & Evolution (2023) | Cite this article

1259 Accesses | 46 Altmetric | Metrics



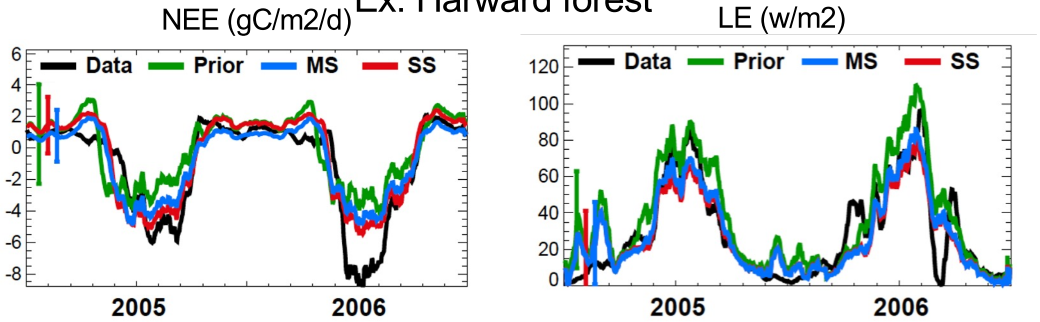
Multiple data streams assimilation

Step 1:
MODIS-NDVI
4 params / PFT

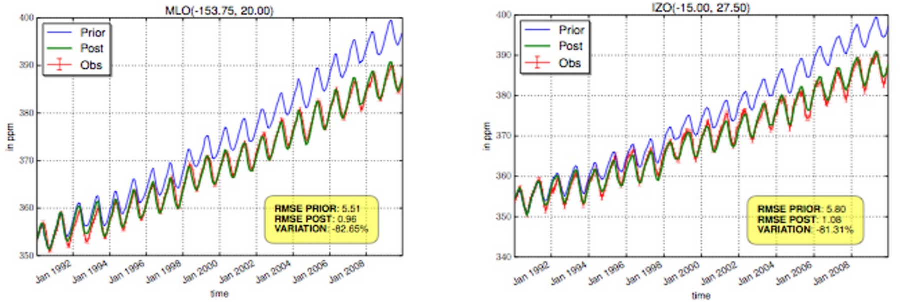


Step 2:
75 fluxnet data
≈ 20 params /PFT

Ex: Harward forest



Step 3:
Atmospheric data
≈ 100 params total

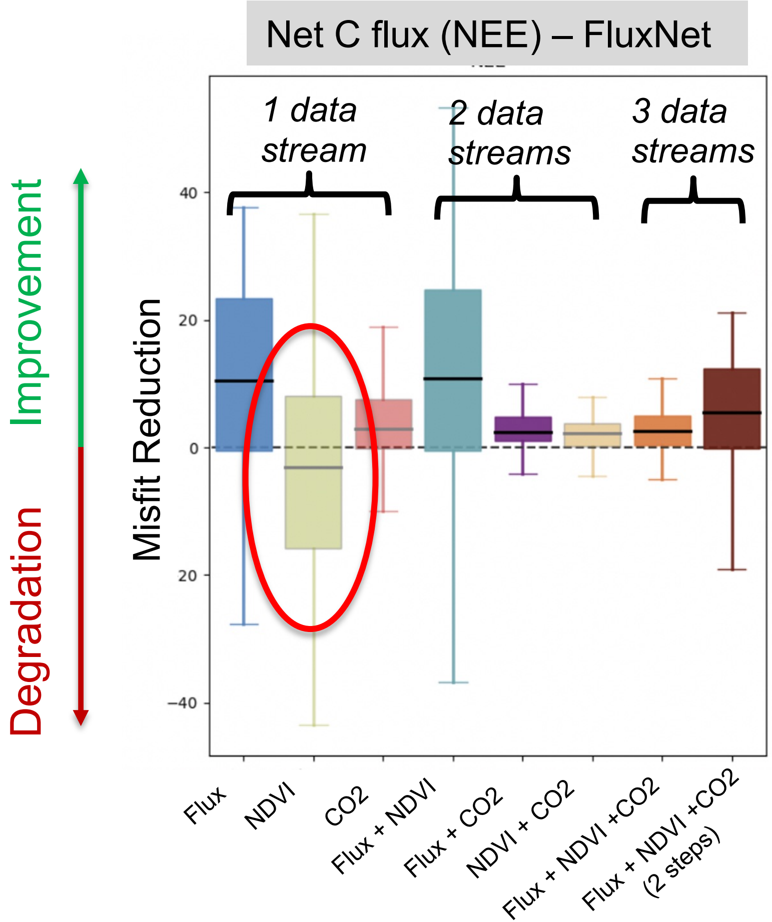


Multiple data streams assimilation

Simultaneous assimilation of all or a few data streams

➔ Different combinations can give different results…

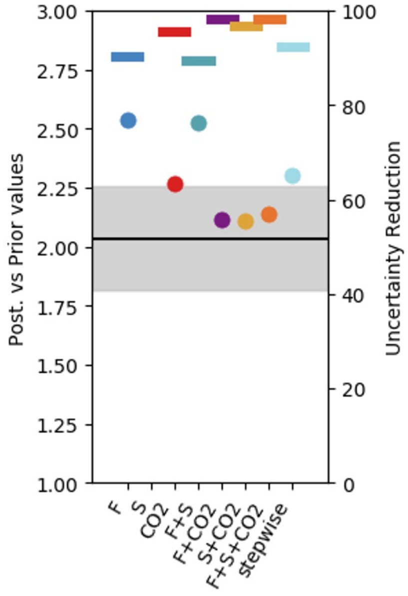
➔ Using only one data stream may degrade the fit to others

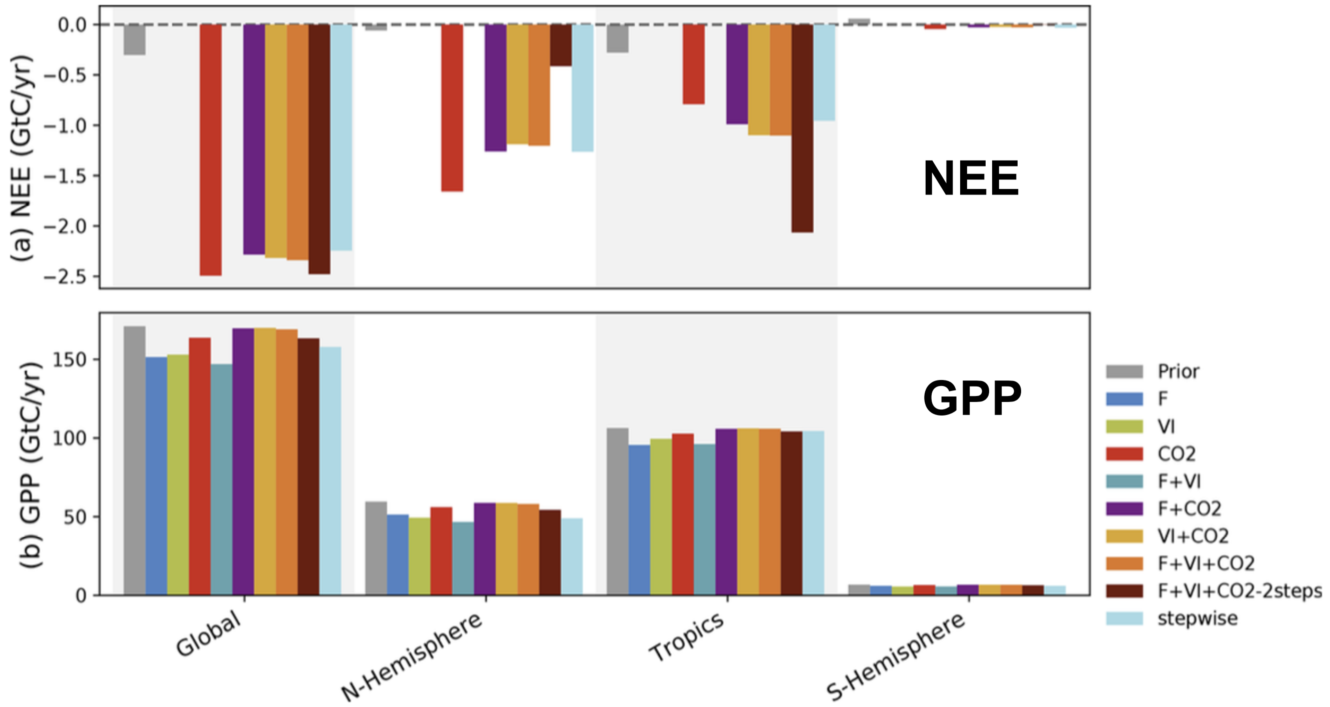


Bacour, C., et al. (2023)

Combining multiple data streams is key to get meaningfull global C fluxes !

Ex. of Q10 param.





➔ Different combinations can give different parameter values

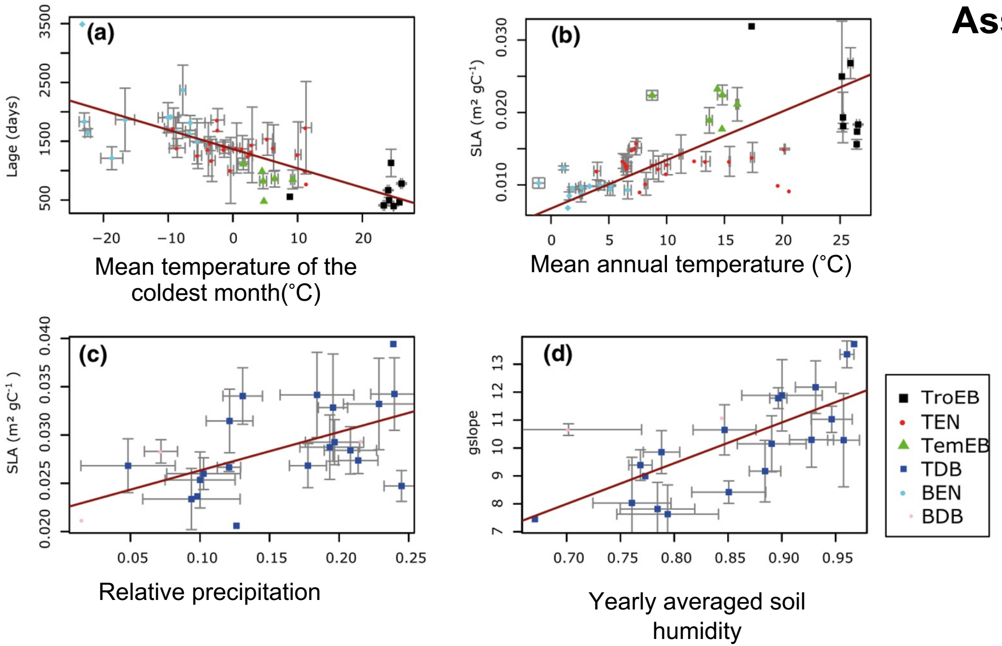
➔ Nee at least Atmospheric CO2 data to have robust global NEE budget !

Bacour et al. (2023)
Biogeosciences

Data Assimilation to highlight ecological relationships

Process understanding

Ecological consistency of optimized trait-related parameters



Assimilation of GPP data / 371 site-years estimates
14 parameters linked to C assimilation

- Optimized parameter values consistent with leaf-scale traits and well-known trade-offs observed at the leaf level
- Sensitivity of trait-related parameters to local bio-climatic variables > **reproduce observed relationships between traits and climate**
- **Indirect validation of the main GPP-related processes implemented in ORCHIDEE**

@Peaucelle et al. (2019)

Assimilation of present-day observations

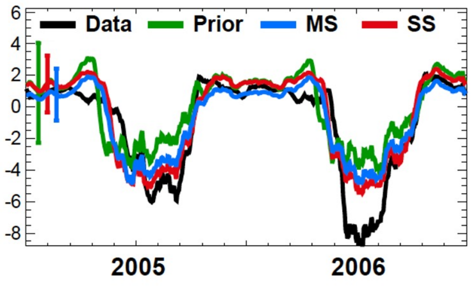
does not guaranty improved future simulations !

Assimilation of present-day observations does not guaranty improved future simulations !

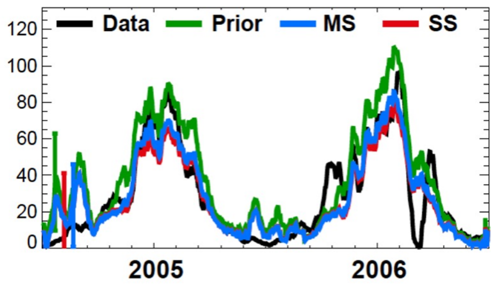
Model improvement when assimilation LE and GPP over FLUXNET broadleaf sites

Ex: Harward forest

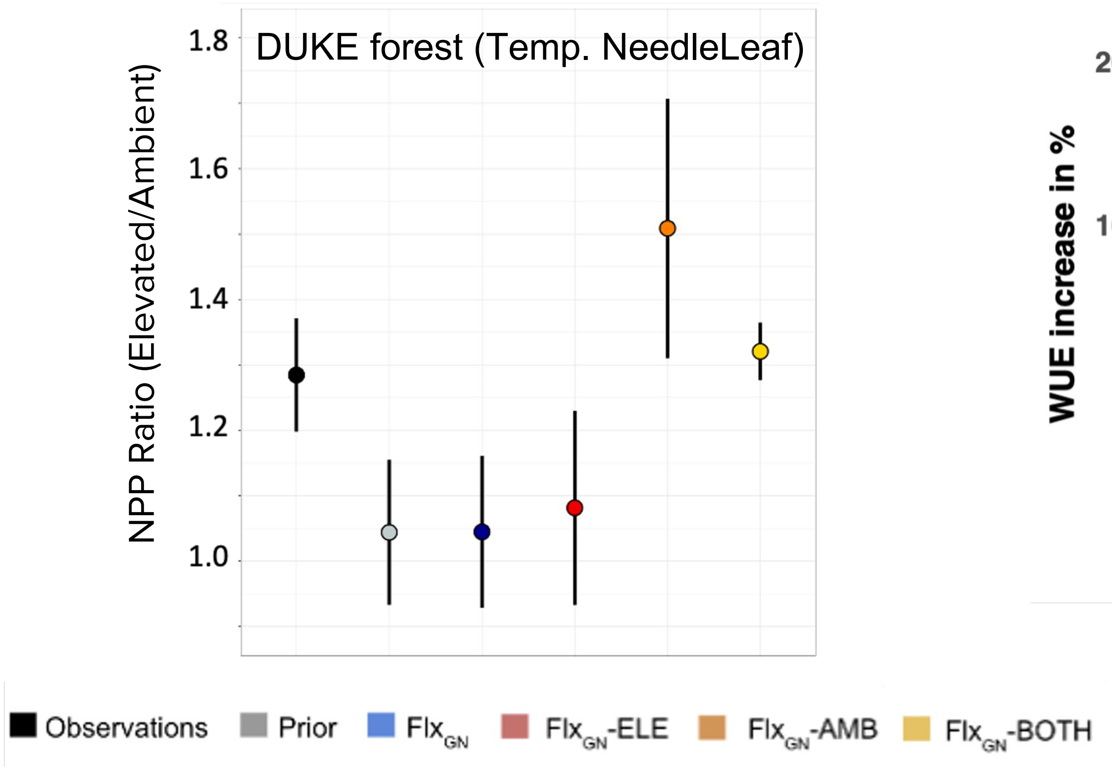
NEE (gC/m2/d)

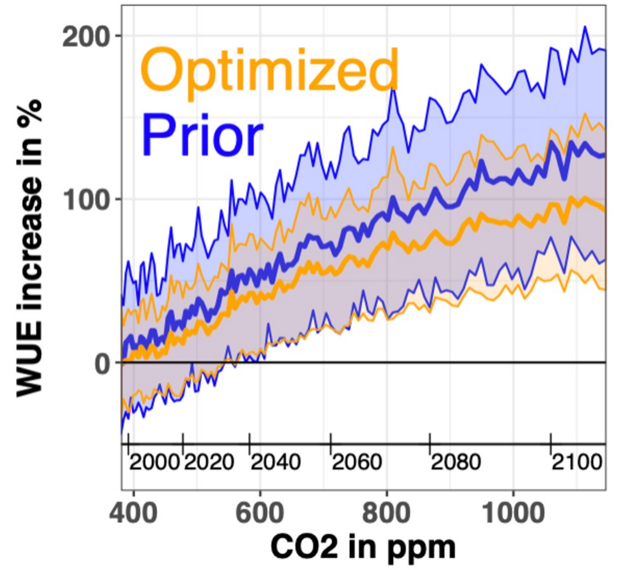


LE (w/m2)



The addition of Free Air CO_2 Enchirement data to the optimisation increases confidence in the optimised model's projections





1% CO_2 increase per year

⇒ ORCHIDEE - CN prior underestimates the change of NPP with doubling CO2

Raoult et al. (submitted)

Outline…

- Motivation for Data Assimilation (parameter calibration)
- Highlights of scientific results and issues linked to parameter optimisation
- **Remaining key challenges & Upcoming opportunities**

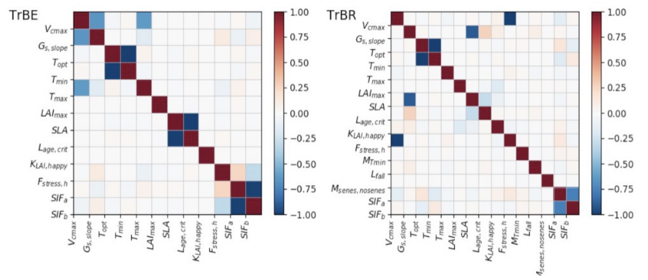
Remaining key challenges

- Model overfitting ➔ degradation of some skills !
  Largely linked to equifinality !

- Including the Spin up in model calibration
  - Crucial for the carbon cycle and soil C pools
  - Difficult because of computing time !

Make use of

Parameter error covariance matrix

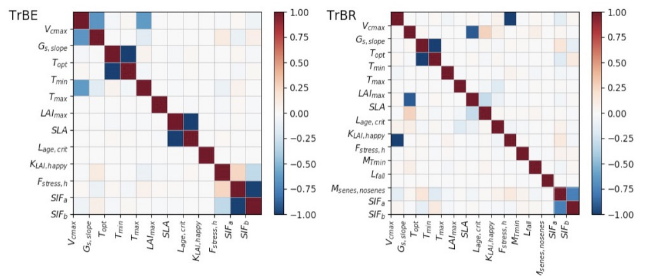


Remaining key challenges

➢ Model overfitting ➔ degradation of some skills !
Largely linked to Overfitting !

Make use of

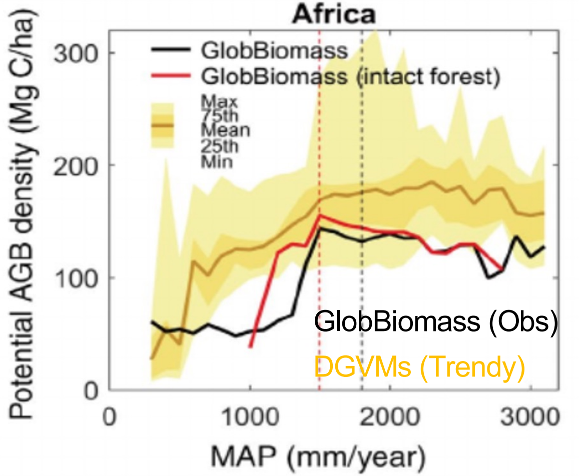
Parameter error covariance matrix



➢ Including the Spin up in model calibration
- Crucial for the carbon cycle and soil C pools
- Difficult because of computing time !

➢ Assimilation of Biomass C pools
rarely attempted but key for the C cycle

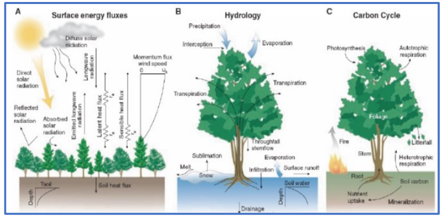
Valoraise satellite data



Yang et al., GCP (2019)

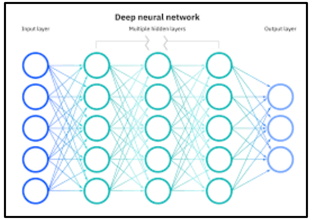
➢ Mixing param and state variable optimisation
to learn on missing processes

Machine Learning algorithms to support parameter optimization



➔ Hybrid modeling :

Process-based vs statistical modeling



Learning the spatial variability of photosynthesis parameters

Shanning Bao[1,2] and Nuno Carvalhais[1]

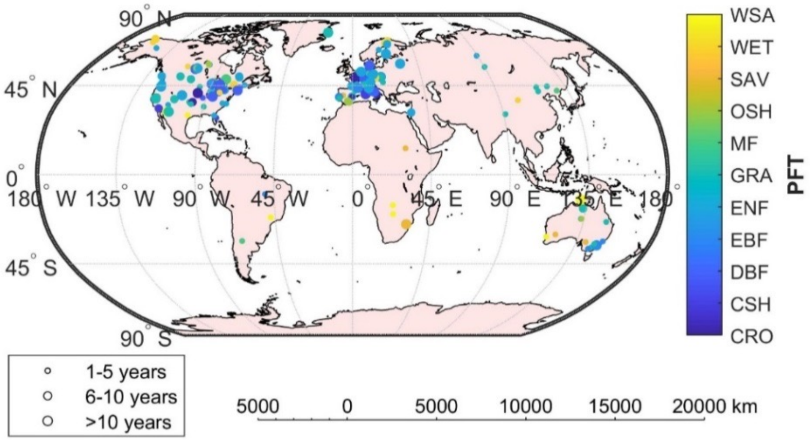
Aim: to predict parameters of photosynthesis model using vegetation + climate + soil properties and constrained by model loss.

Max Planck Institute
for Biogeochemistry


[Bao et al., 2023, JAMES]

$$\mathrm{GPP} = \boldsymbol{\varepsilon_{max}} \cdot \mathrm{PAR} \cdot \mathrm{FAPAR} \cdot f\mathrm{T} \cdot f\mathrm{VPD} \cdot f\mathrm{W} \cdot f\mathrm{L} \cdot f\mathrm{CI}$$

- Semi-empirical descriptions of “ *f* ”
  - Sensitivity of ecosystem GPP to different forcing (climate, soil,..)

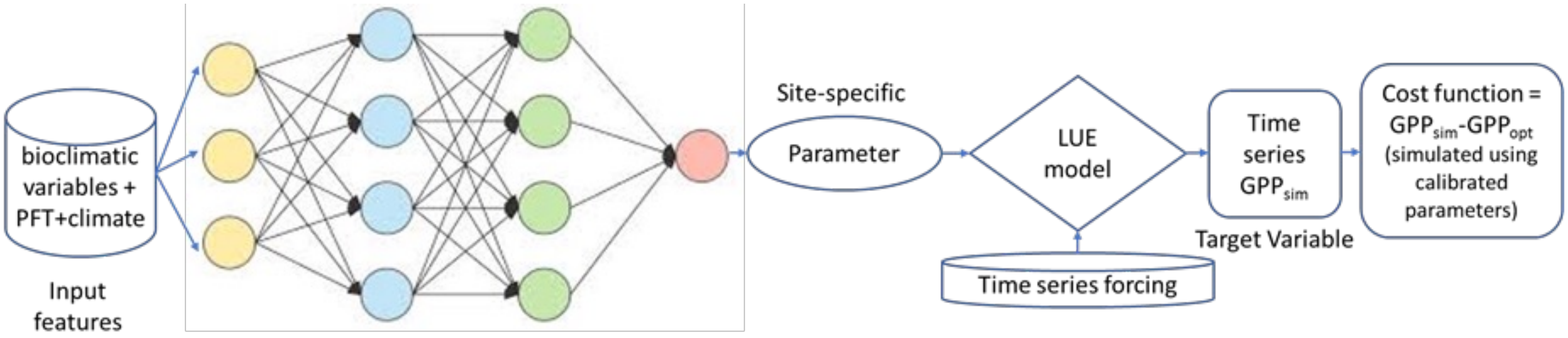


[Bao et al., 2023, JAMES]

$$\mathrm{GPP} = \varepsilon_{max} \cdot \mathrm{PAR} \cdot \mathrm{FAPAR} \cdot f\mathrm{T} \cdot f\mathrm{VPD} \cdot f\mathrm{W} \cdot f\mathrm{L} \cdot f\mathrm{CI}$$

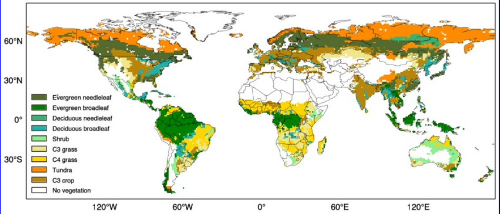
- Semi-empirical descriptions of " *f* "
  - Sensitivity of ecosystem GPP to different forcing (climate, soil,..)

➔ **Learn (NN) the spatial distribution of photosynthesis parameters from FLUXNET GPP observations**



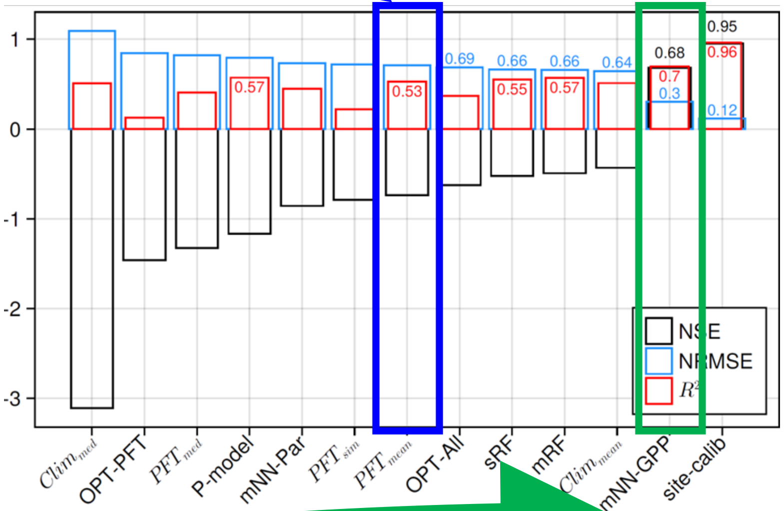
[Bao et al., 2023, JAMES]

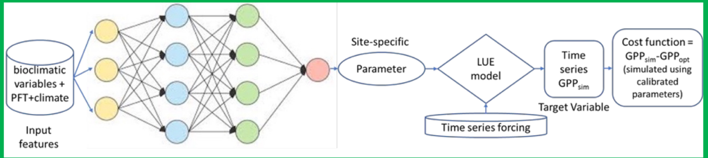
$$GPP = \varepsilon_{max} \cdot PAR \cdot FAPAR \cdot fT \cdot fVPD \cdot fW \cdot fL \cdot fCI$$

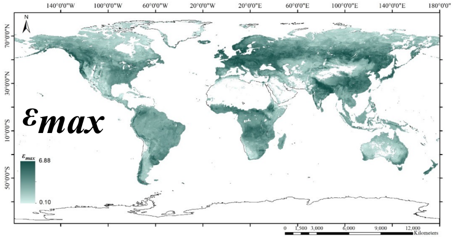


Classical PFT approach

Evaluation of model performances (at Fluxnet sites)







Maps of param. distributions

[Bao et al., 2023, JAMES]

New DA methods emerge with the use of emulators !

➔ History Matching provides an alternative Bayesian approach to model calibration

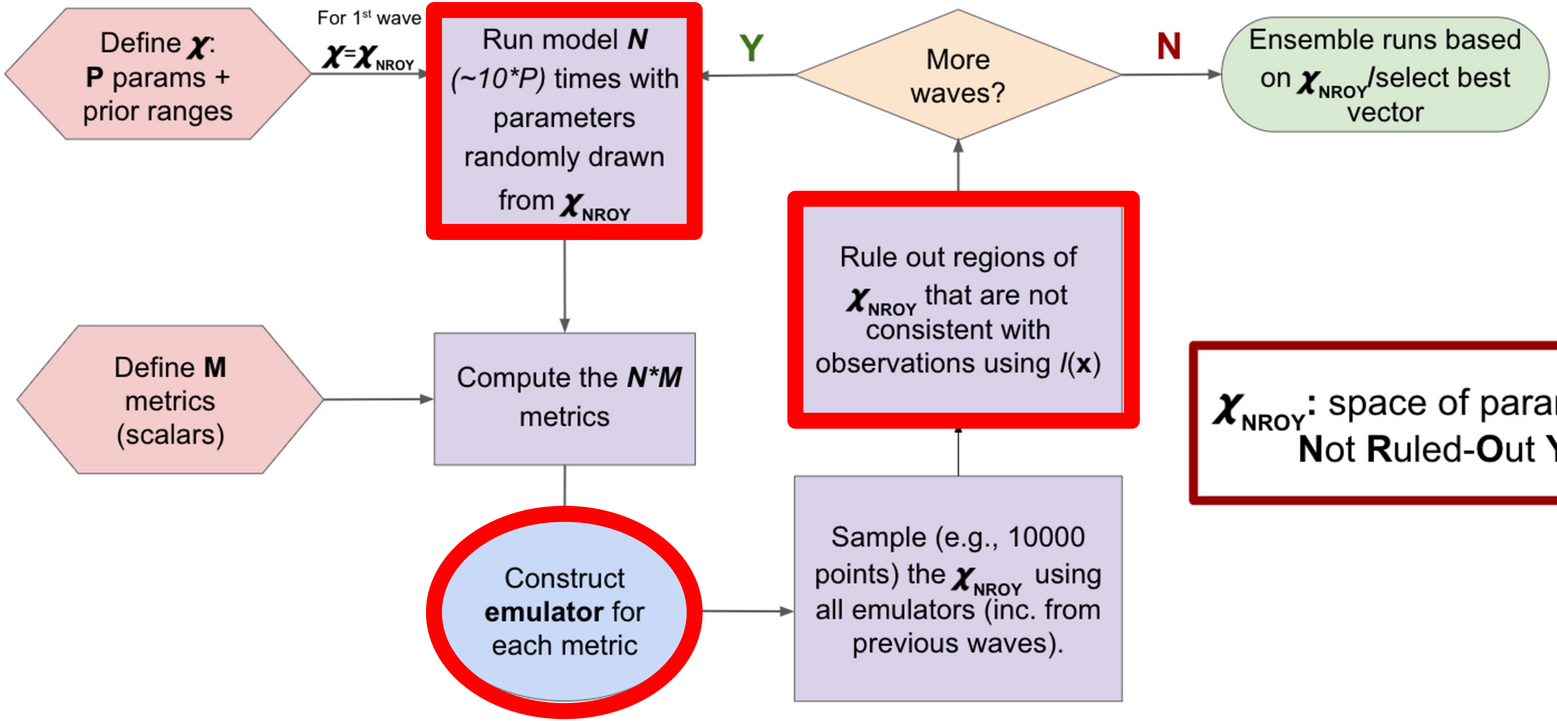
$$J(\mathbf{x}) = \frac{1}{2}\left[(H(\mathbf{x}) - \mathbf{z})^T \mathbf{R}^{-1} (H(\mathbf{x}) - \mathbf{z}) + (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b)\right].$$

Bayesian Calibration
Find likely parameters minimising a cost function

$$\mathcal{I}(\mathbf{x}) = \frac{|\mathbf{z} - \mathrm{E}[H(\mathbf{x})]|}{\sqrt{\mathrm{Var}[\mathbf{z} - E[H(\mathbf{x})]]}} = \frac{|\mathbf{z} - \mathrm{E}[H(\mathbf{x})]|}{\sqrt{\mathrm{Var}[H(\mathbf{x})] + \mathrm{Var}[e] + \mathrm{Var}[\eta]}}.$$

History Matching
Rule out unlikely parameters using an implausibility function

History matching: based on Gaussian Emulators



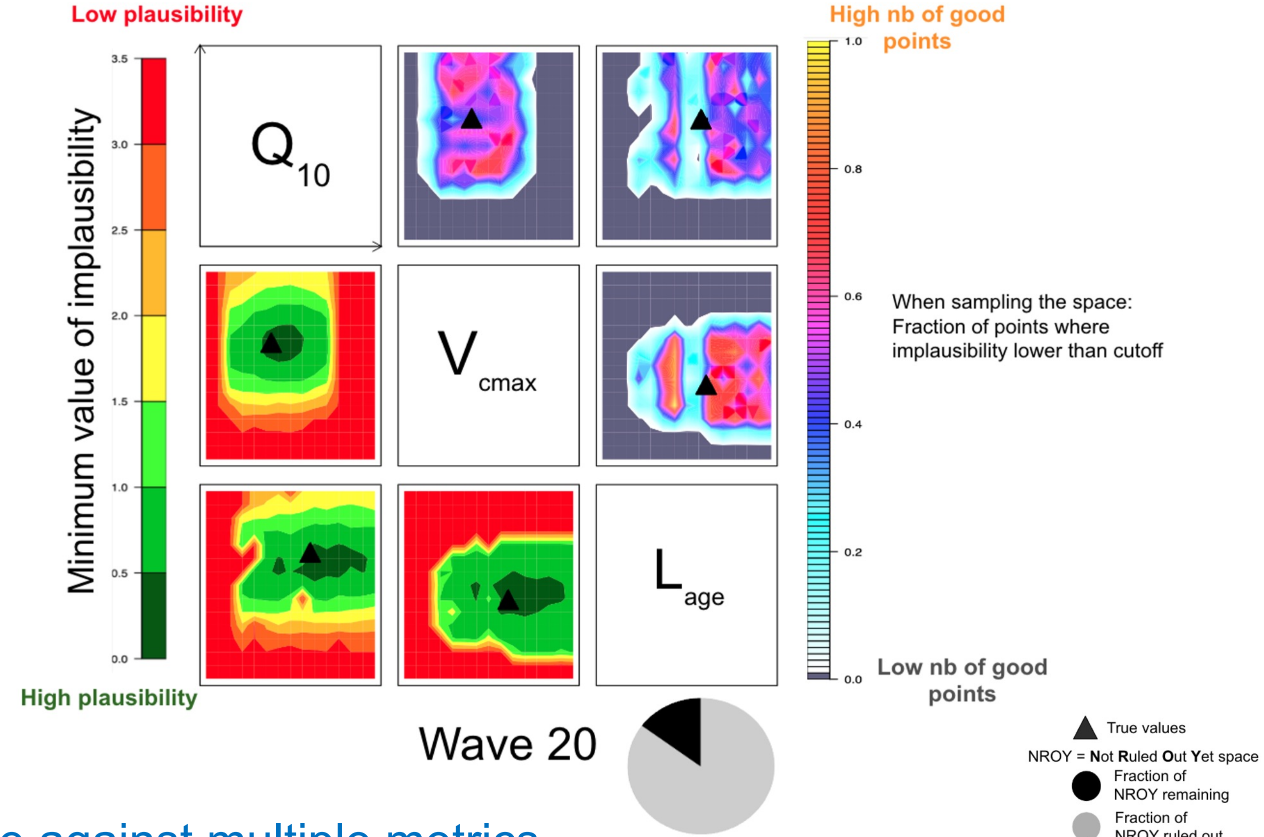
$\boldsymbol{\chi}_{NROY}$: space of parameters **N**ot **R**uled-**O**ut **Y**et

$$\mathcal{X}_{NROY} = \{\mathbf{x} \in \mathcal{X} | \mathcal{I}(\mathbf{x}) < 3^2\}$$

History matching

Twin experiment (known param values):

- After 20 waves removed 7/8 of space as highly unlikely
- True values are in highly plausible space



➔ Advantages: ability to tune against multiple metrics

➔ Easy to generate ensemble from posterior distribution

Summary and key issues

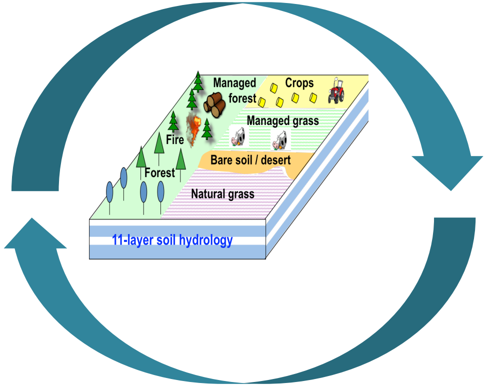
- Data assimilation (parameter optimisation) should play a big role in reducing uncertainties and inter model spread in model predictions of C / W / E fluxes

- However, model structural error is a critical issue with parameter optimisation ! And "**overfitting**" often breaks the overall model skills !

- Available in situ and satellite data are still largely under-used to calibrate global land surface models

- Model improvement should be a central part of the process! when optimisation fail to fit the observations ➔ Highlight structural errors or forcing errors !

Model development



Parameter optimisation

Thank you ….

➔ Welcome to join international initiatives on Data Assimilation

https://hydro-jules.org/international-land-modeling-forum-ilmf?utm_source

https://land-da-community.github.io

https://aimesproject.org/ldawg/

International Land Modelling Forum (ILMF) Webinars

Save the date! The dates and times are listed below with a meeting invite for you to save it in your calendar.

International Land Modelling Forum (ILMF) Interactive Webinars
September – October 2023

The ILMF is kicking off a set of initial Working Groups through a series of 2–3-hour interactive webinars. Conveners of the webinars will be arranging speakers and leading a discussion around the goals and organization of each Working Group. All are welcome. Please join us! QR codes are clickable.

10th October 2023: 14:00 (UTC)
Parameter Estimation Methods for Land Models
Rosie Fisher, Nina Raoult and Daniel Kennedy

Register here for Parameter Estimation Methods for Land Models

18th October 2023: 14:00 (UTC)
Incorporating Human Activities More Comprehensively into Land Models
Sonali McDermid, Danica Lombardozzi and Julia Pongratz

Register here for Incorporating Human Activities More Comprehensively into Land Models

Land Data Assimilation Community | About | Join! | Events | News and Opportunities | Publications | Training

Land DA Community
Welcome to the Land DA Community Website!

Email
Github

Welcome to the Land DA Community Website!

This website will serve as a hub for all Land DA Community activities, resources, and announcements.

Please check out the pages above to see past and planned events, DA tutorials, land DA-related publications, job adverts, and more! You can also join the land DA community email listserv by clicking on the "Join!" tab above. Let us know via email if you have any suggestions for how to improve this website.

– This website is maintained by the AIMES Land DA Working Group. Find out more here.

Additional slides

New methods

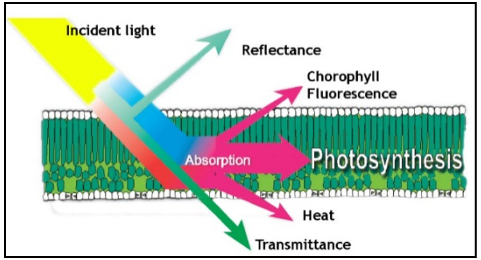
The Land Variational Ensemble Data Assimilation Framework: LAVENDAR v1.0.0

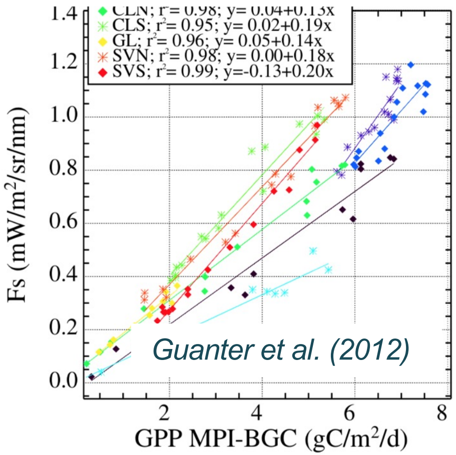
Ewan Pinnington[1], Tristan Quaife[1,2], Amos Lawless[1,2], Karina Williams[3], Tim Arkebauer[4], and Dave Scoby[4]

- Essentially 4DVar without needing an adjoint or TLM
- Ensemble generation and analysis are completely separate
- We typical use 20-50 ensemble members (can be slow), depending on problem
- **But** analysis step is *extremely fast*
  - Don't need to run the model!
  - 9M observations in a few minutes for Africa example
- Consequently, once an ensemble is built it is possible to run multiple experiments with it (to examine the impact of different observations)
- https://github.com/tquaife/4DEnVar_engine

New data are coming with associated challenges !

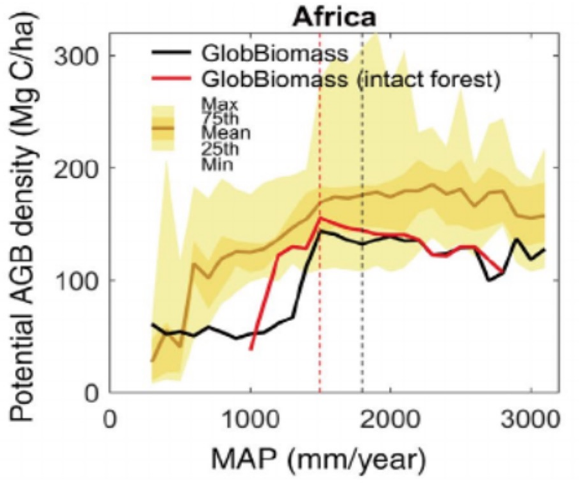
Solar Induced fluorescence (SIF)





Guanter et al. (2012)

Satellite biomass data

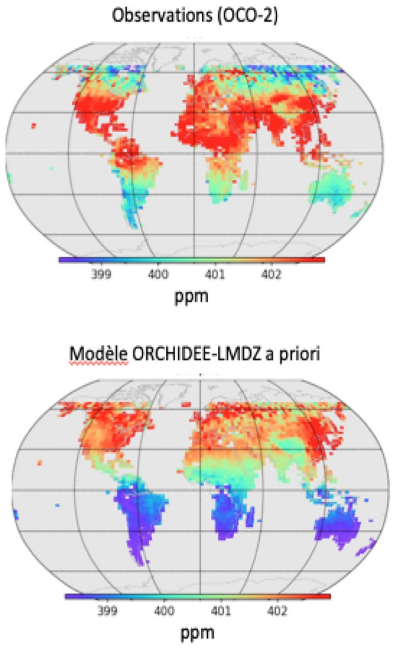


GlobBiomass (Obs)

DGVMs (Trendy)

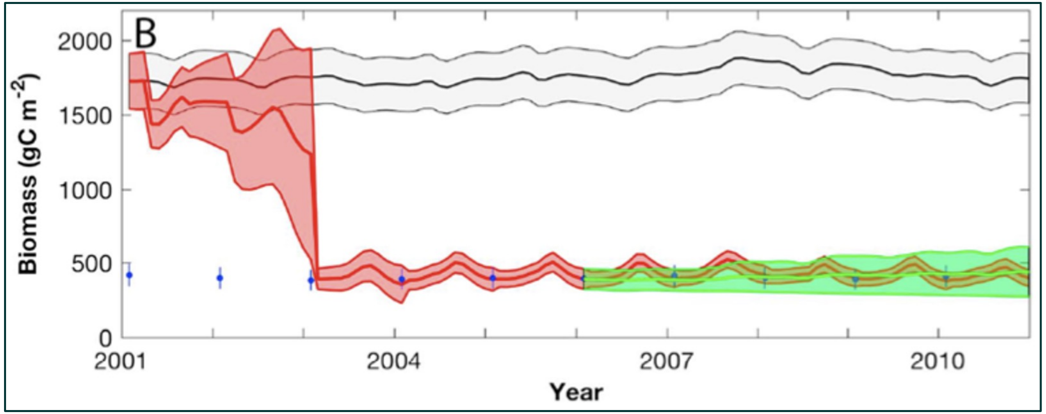
Yang et al., GCP (2019)

Satellite XCO2 data



On-Going work

State Data Assimilation for updating C stocks and fluxes with CLM



Evaluation of a Data Assimilation System for Land Surface Models Using CLM4.5

Andrew M. Fox[1], Timothy J. Hoar[2], Jeffrey L. Anderson[2], Avelino F. Arellano[3], William K. Smith[1], Marcy E. Litvak[4], Natasha MacBean[1], David S. Schimel[5], and David J. P. Moore[1]

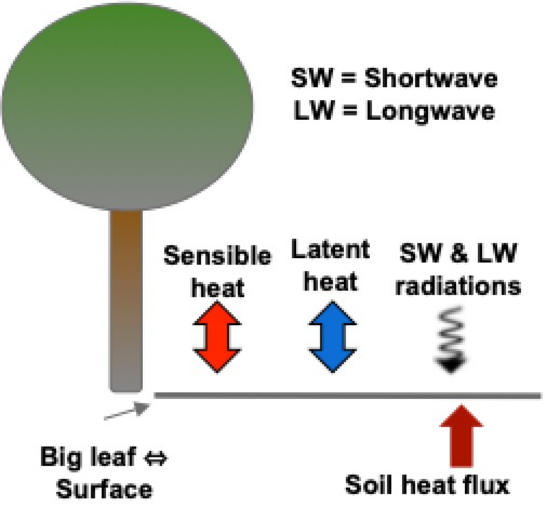
Assimilation of Global Satellite Leaf Area Estimates Reduces Modeled Global Carbon Uptake and Energy Loss by Terrestrial Ecosystems

Andrew M. Fox[1], Xueli Huo[2], Timothy J. Hoar[3], Hamid Dashti[2], William K. Smith[2], Natasha MacBean[4], Jeffrey L. Anderson[3], Matthew Roby[2], and David J. P. Moore[2]

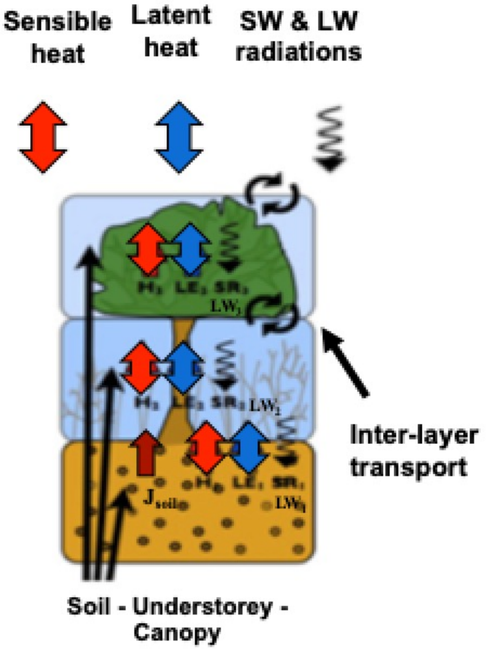
Improving CLM5.0 Biomass and Carbon Exchange Across the Western United States Using a Data Assimilation System

Brett Raczka[1,2], Timothy J. Hoar[3], Henrique F. Duarte[4,5], Andrew M. Fox[6], Jeffrey L. Anderson[3], David R. Bowling[1,4], and John C. Lin[4]

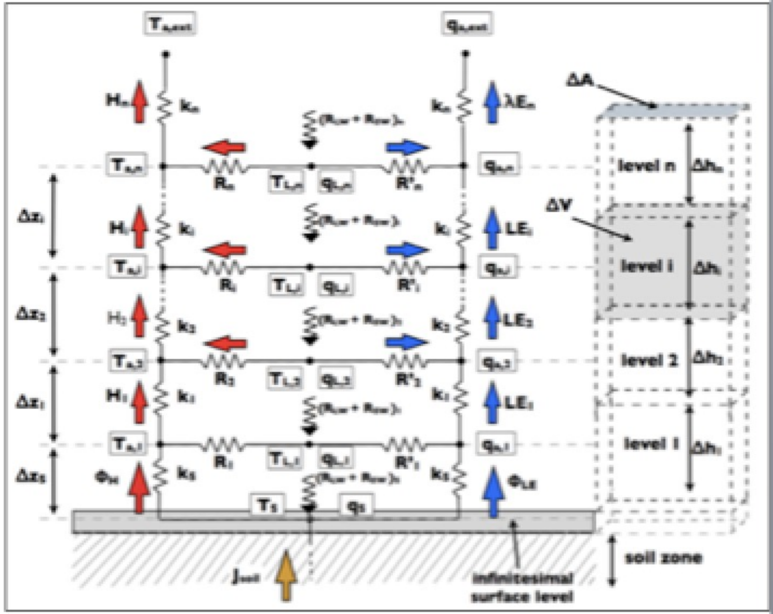
“Big leaf” model



Two-layer energy budget



Multi-layer energy budget

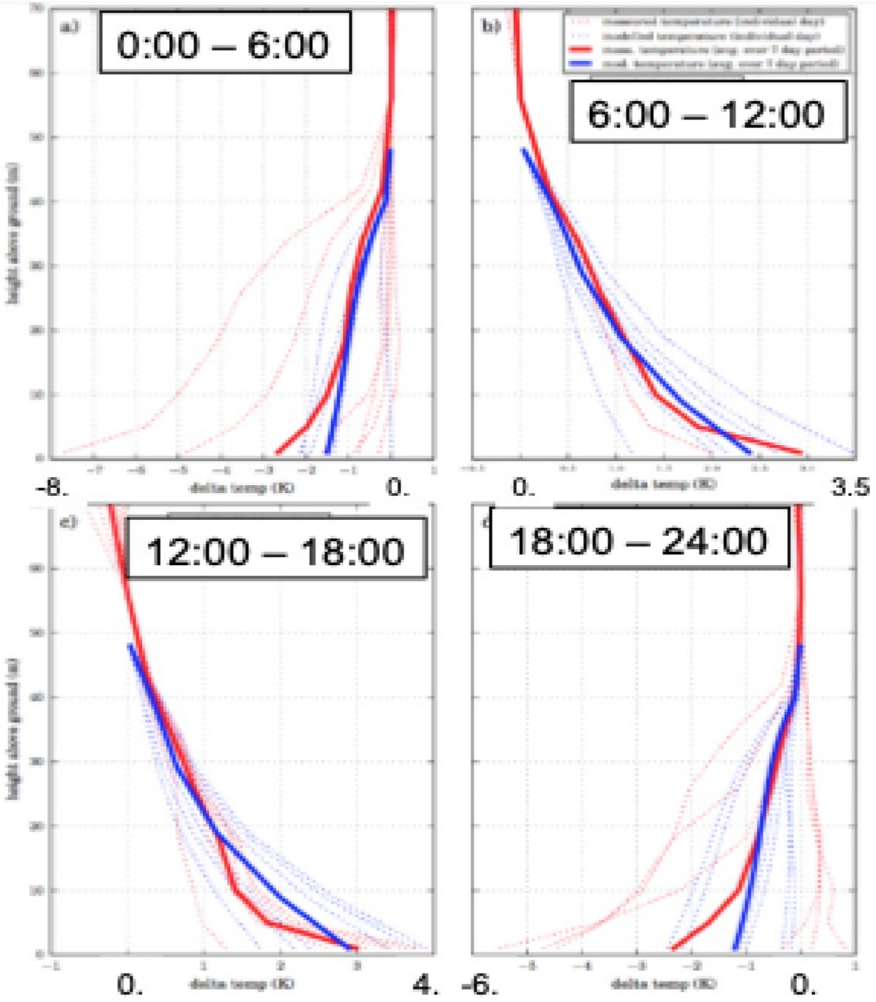


Turbulent transfer of heat

Turbulent transfer of water vapor

Radiation transferts

Complexity



Daily temperature

Temperature profile at Tumbarumba site

Observations

Model

Ryder et al., 2015

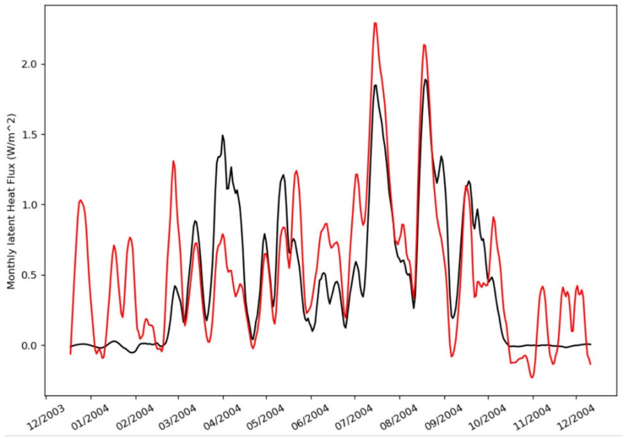
Recent results in the Trunk of ORCHIDEE

Multi-layer energy budget: Local results (DE-Hai – 2004)

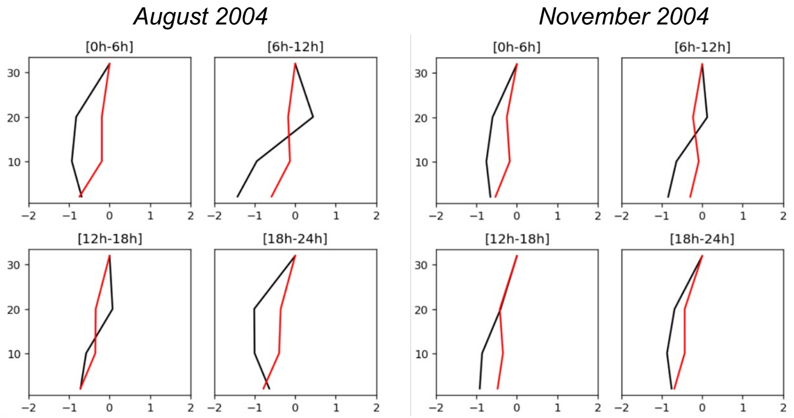
Multi-Layer EB

Observations

Temperature difference between top canopy and surface in 2004



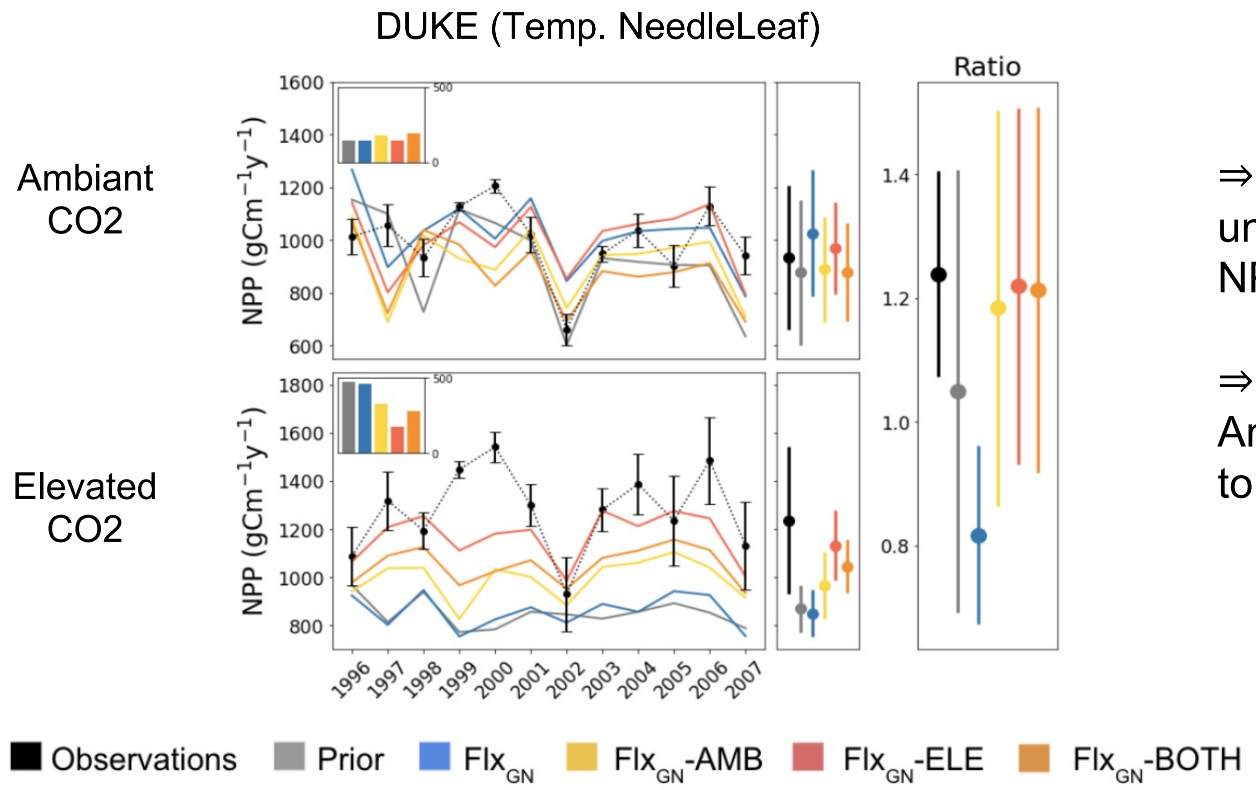
Normalized intra-canopy temperature gradient



- **Overall canopy temperature gradient dynamics** well represented during the year;
- **Intra-canopy climate well reproduced** most of the time;

Assimilation of Free Air CO2 Enrichment data (FACE)

➔ Optimisation of ORCHIDEE params (~ 20) at FACE sites (Oak Rige & Duke) with NPP & LAI



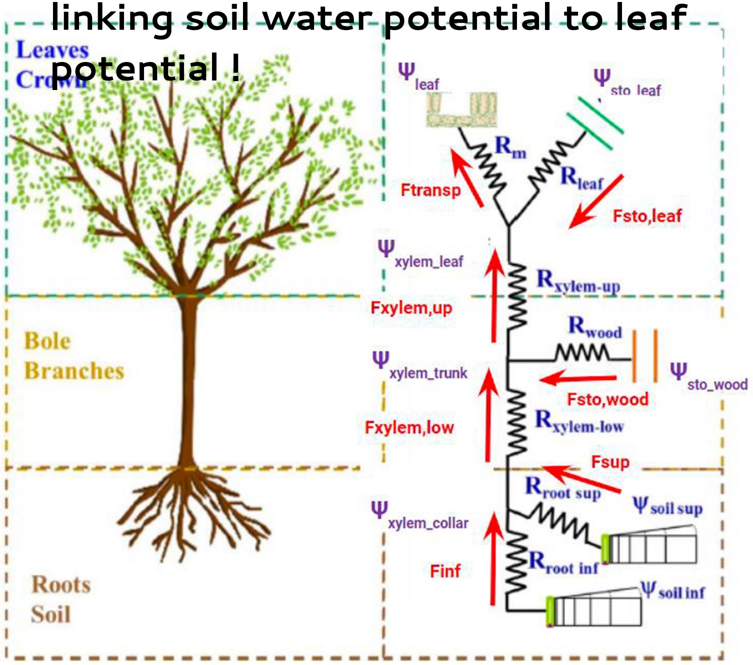
⇒ ORCHIDEE - CN Prior underestimates the change of NPP with doubling CO2

⇒ Need to optimise against both Ambient and Elevated CO2 data to fit the observed NPP ratio

Raoult., et al. (in prep)

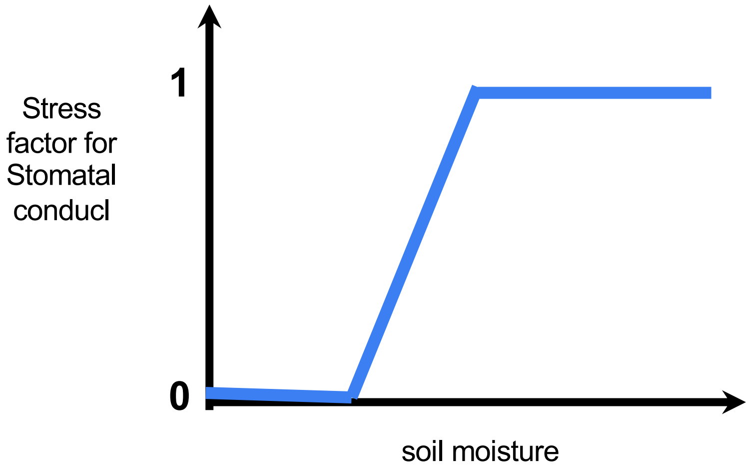
Optimisation of new hydraulic architecture

- **Implementation of a new physical scheme linking soil water potential to leaf potential !**



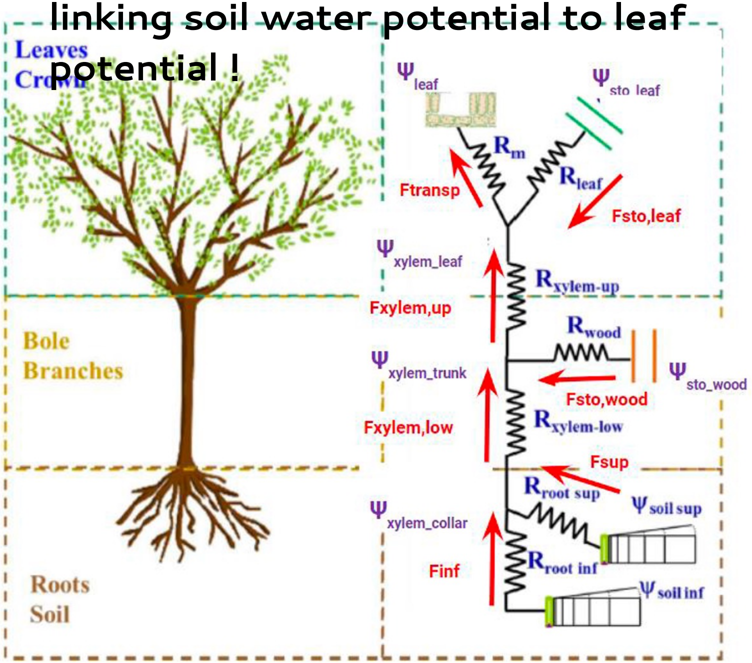
@Alleon et al. (in

Versus a standard statistical scheme to link leaf transpiration / GPP to soil moisture stress



Optimisation of new hydraulic architecture

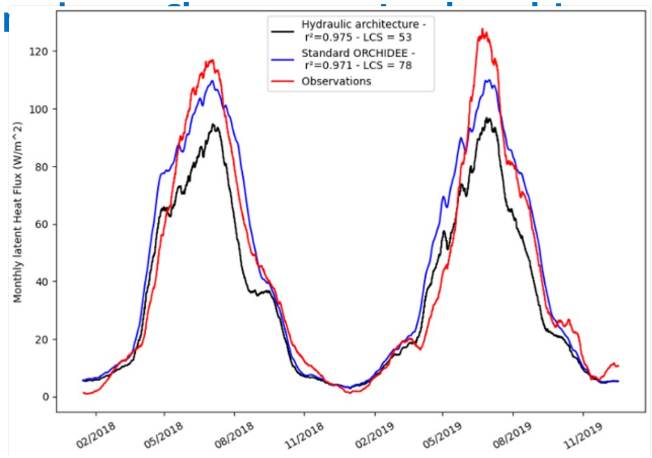
- **Implementation of a new physical scheme linking soil water potential to leaf potential !**



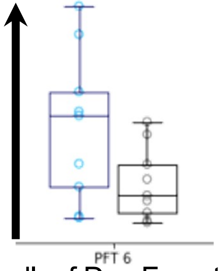
@Alleon et al. (in

- **Optimisation of the STD vs NEW scheme with FluxNet laten**

Hydraulic architecture: r^2=0.975 / LCS=53
Standard ORCHIDEE: r^2=0.971 / LCS=78
Observations



==> Higher capability to model temporal flux variations especially during droughts !



Broadleaf Dec Forest