Data assimilation with ORCHIDEE

Natasha MacBean, Philippe Peylin, Cédric Bacour, Sebastien Leonard, Fabienne Maignan, Philippe Ciais

Laboratoire des Sciences du Climat et de l'Environnement, France

Outline

- What is Data Assimilation?
- Why do we need DA?
- Example 1: Optimising the phenology of ORCHIDEE
- Example 2: Multi-site optimisation with FluxNet
- Example 3: Optimising the phenology with multiple data streams
- Example 4: DA Inter-comparison study

What is data assimilation (DA)?

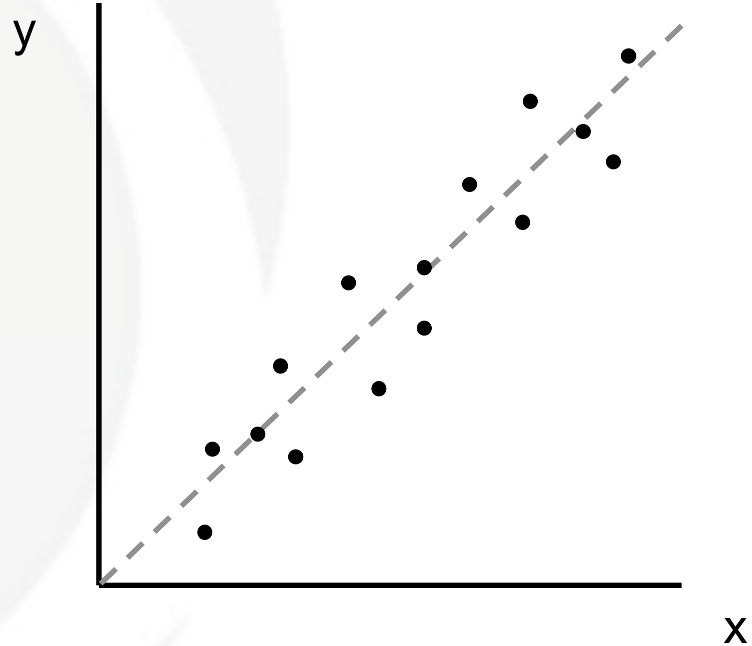
- Also referred to as "Model-Data Fusion" (MDF)…
- Data assimilation comprises of a set of statistical techniques aimed at integrating models (prior knowledge of a system) and observations (new information) to improve model predictions and to obtain an estimate of the distribution of the model prediction (i.e. the uncertainty)
- Based on Bayes' Theorem → update the prior probability of a hypothesis given new observations or evidence
- Basis of DA → the process of combining data with prior knowledge of the variables of a physical system to obtain an improved estimate of the variables

P(model, given the data) α P(model) x P(observations given the model)

What is data assimilation (DA)?

- Can optimise model state variables, initial conditions or parameters
- Here we're talking about parameter (and initial condition) optimisation
- Describe the misfit between the observations and the model simulations, *accounting for the uncertainty in both*
- Try to MINIMIZE the misfit

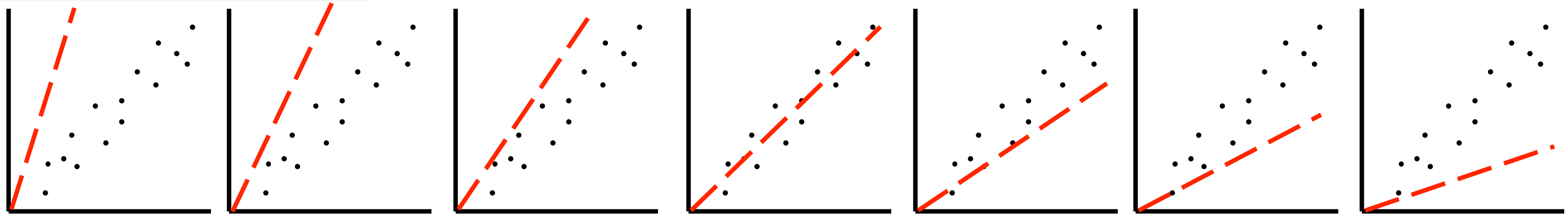
Let's take the simplest case…



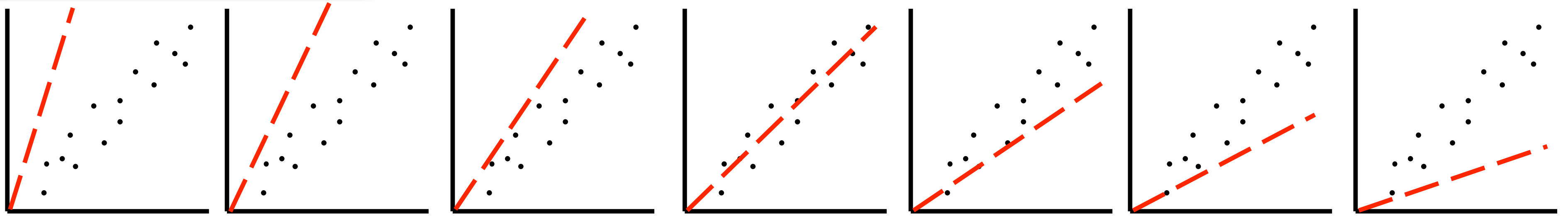
$$y = ax$$

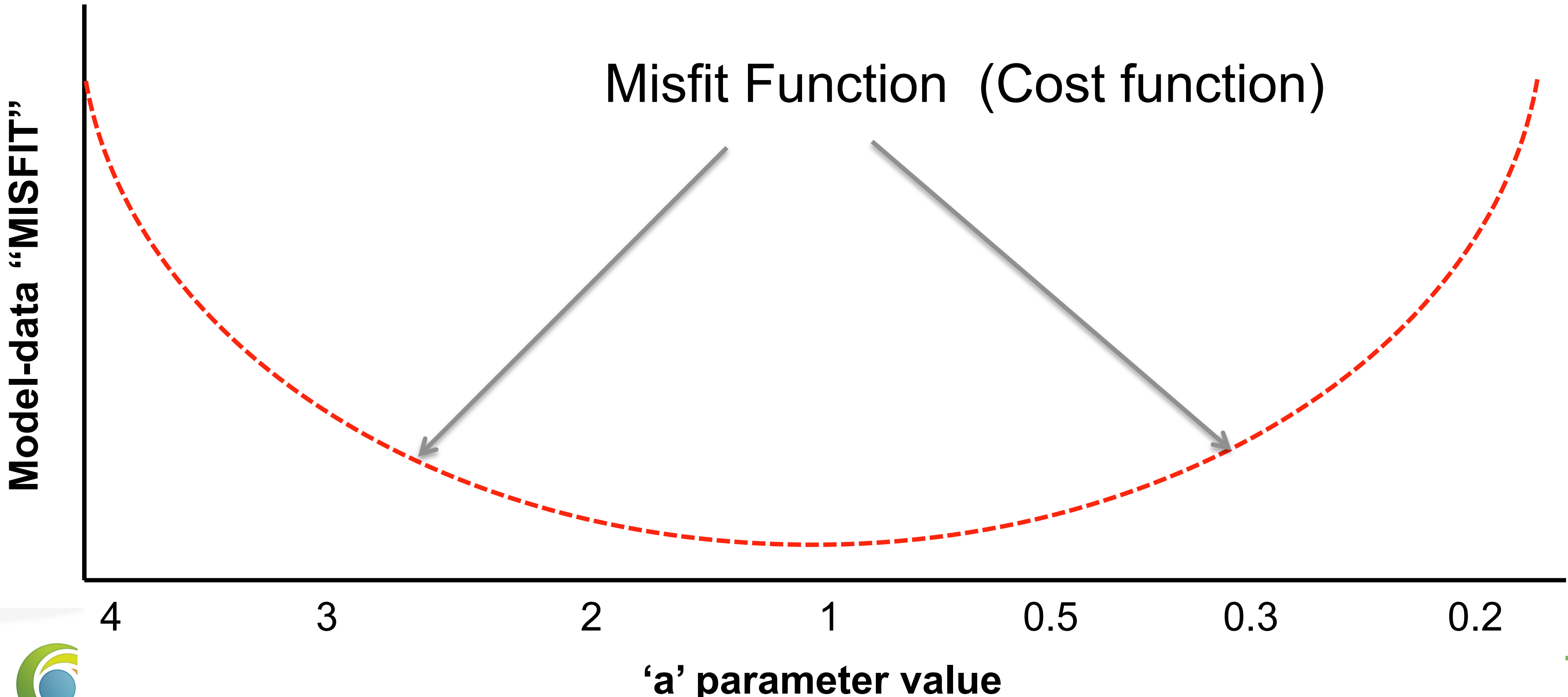
Try to estimate parameter 'a'

Data assimilation for Dummies!

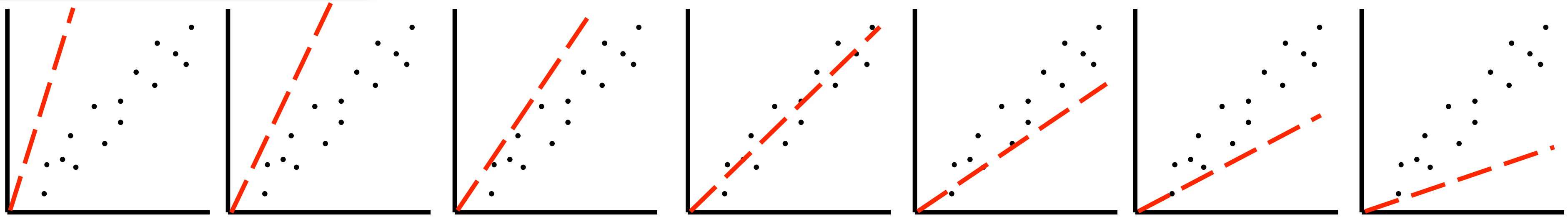


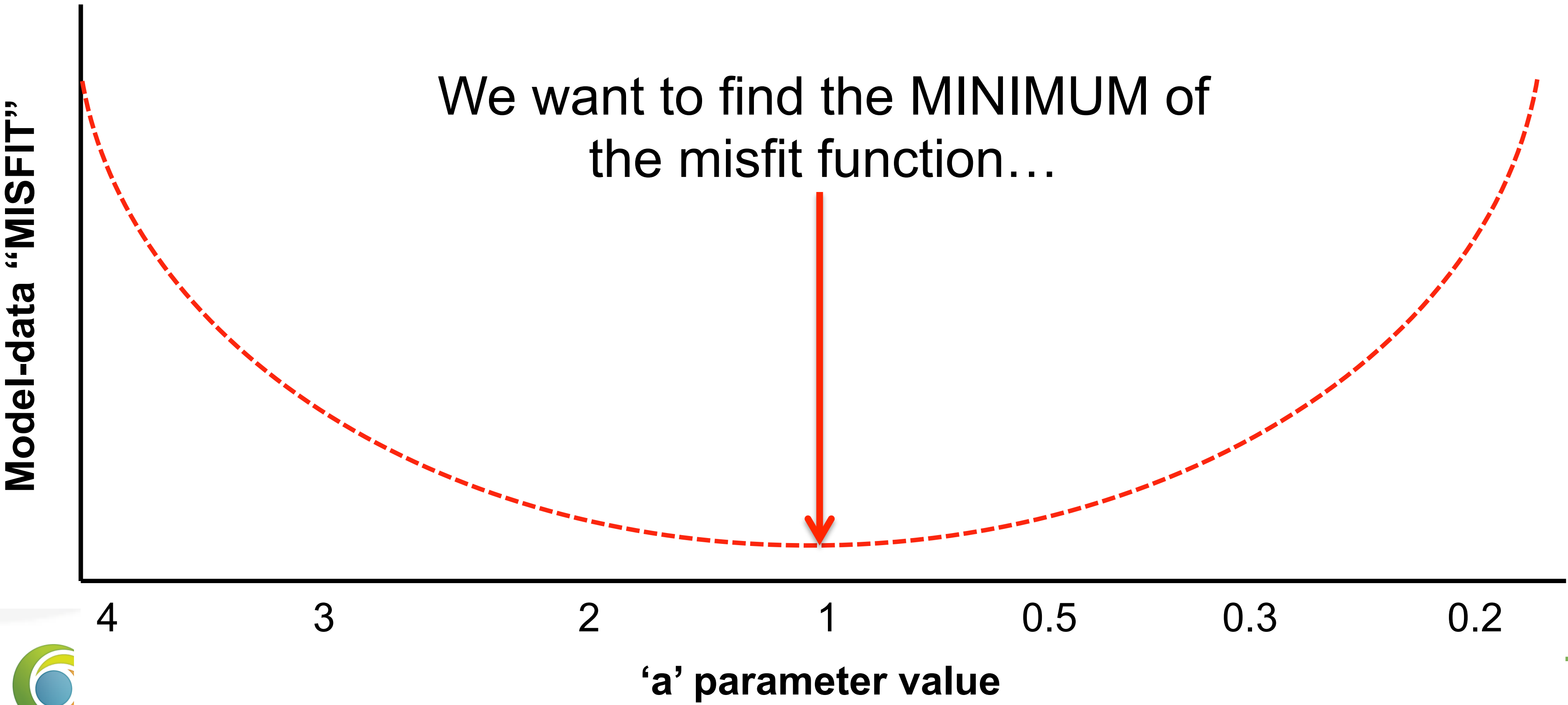
Simplest case!





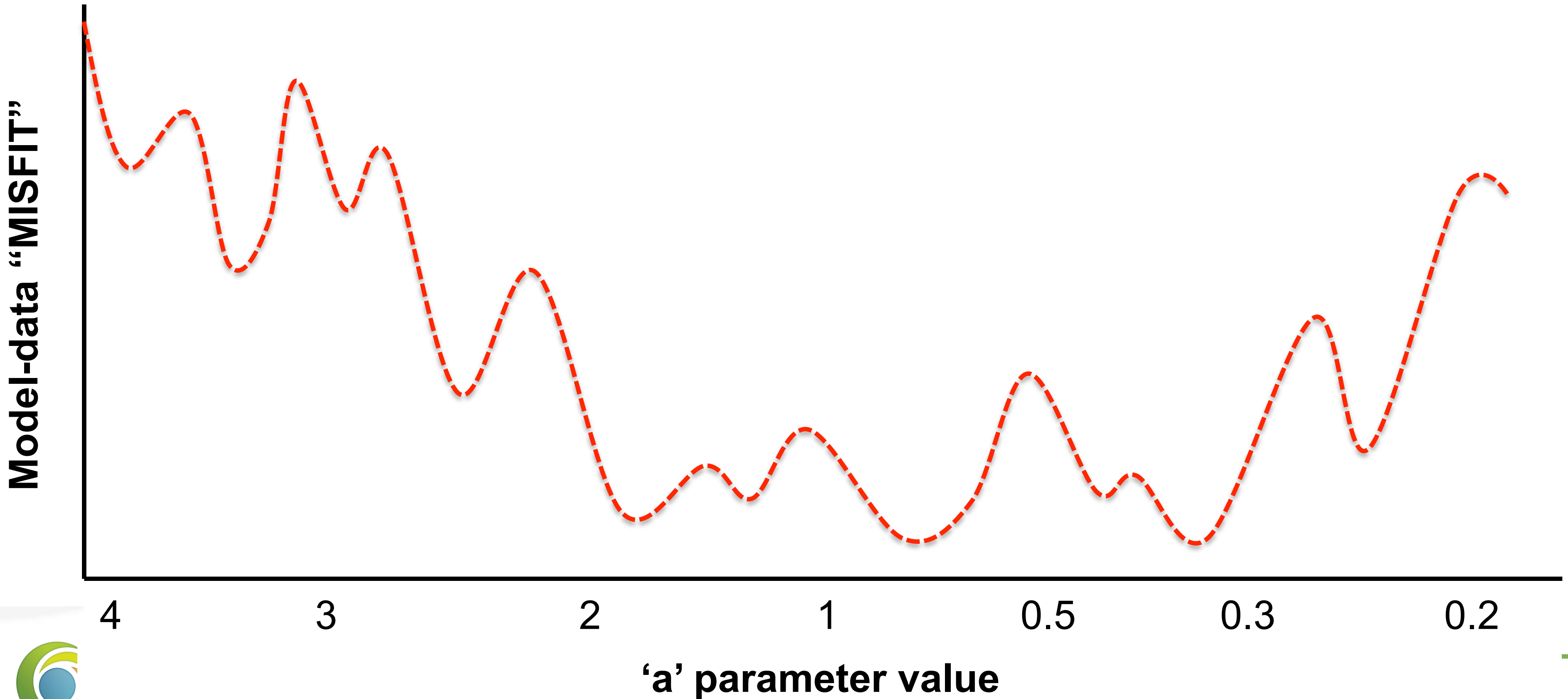
Simplest case!





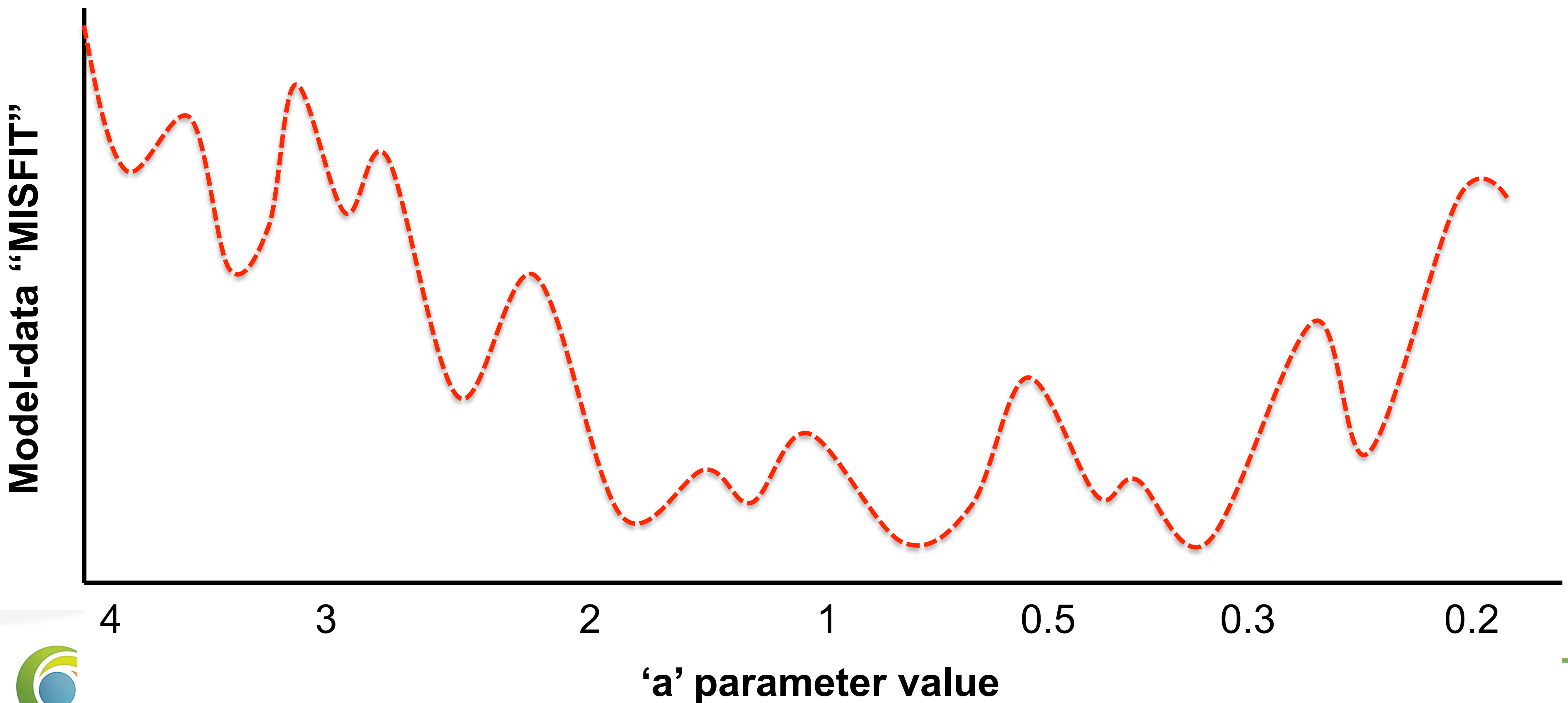
Not so simple case!

- We want to find the MINIMUM of the misfit function…
- BUT! Your misfit function may look like this…!!



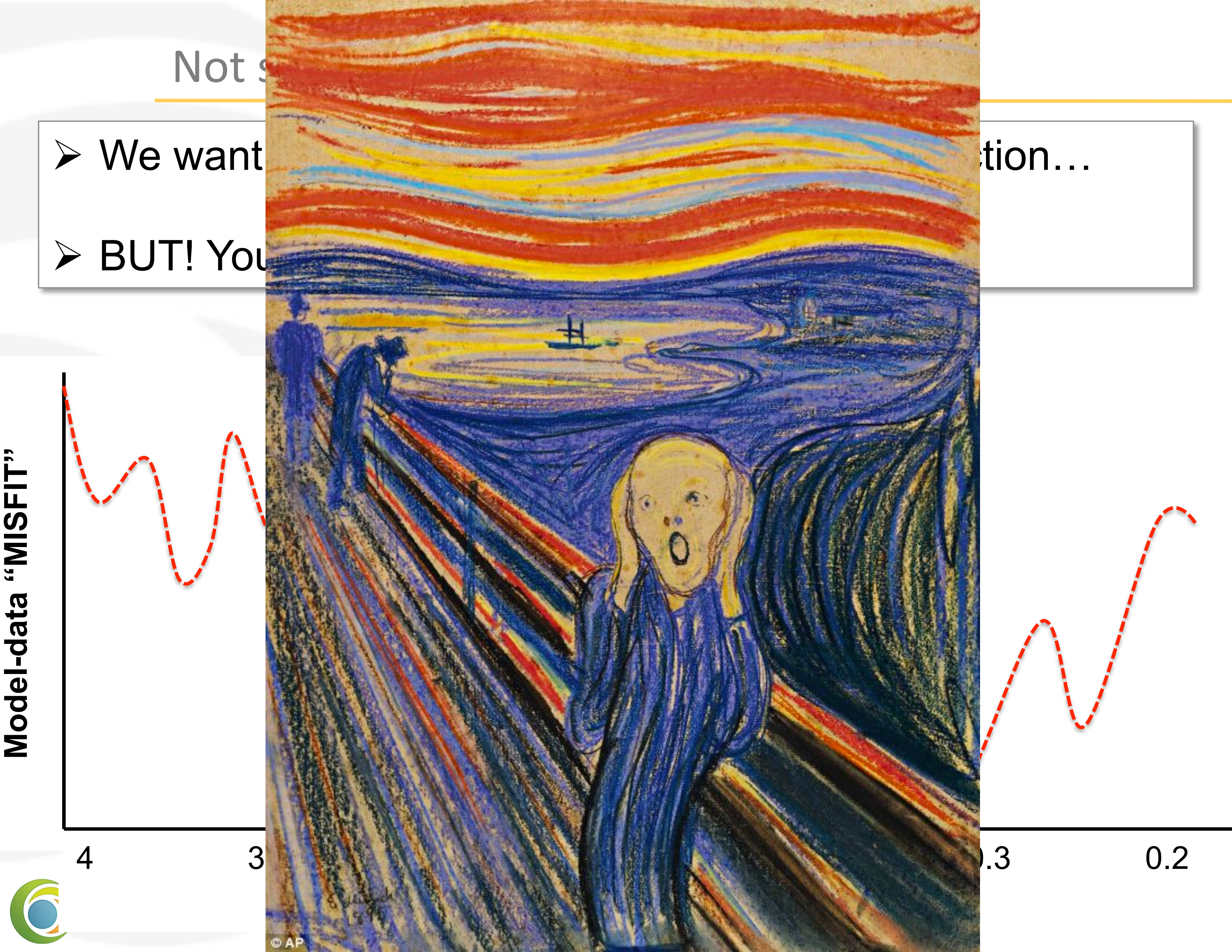
Not so simple case!

- We want to find the MINIMUM of the misfit function…
- BUT! Your misfit function may look like this…!!



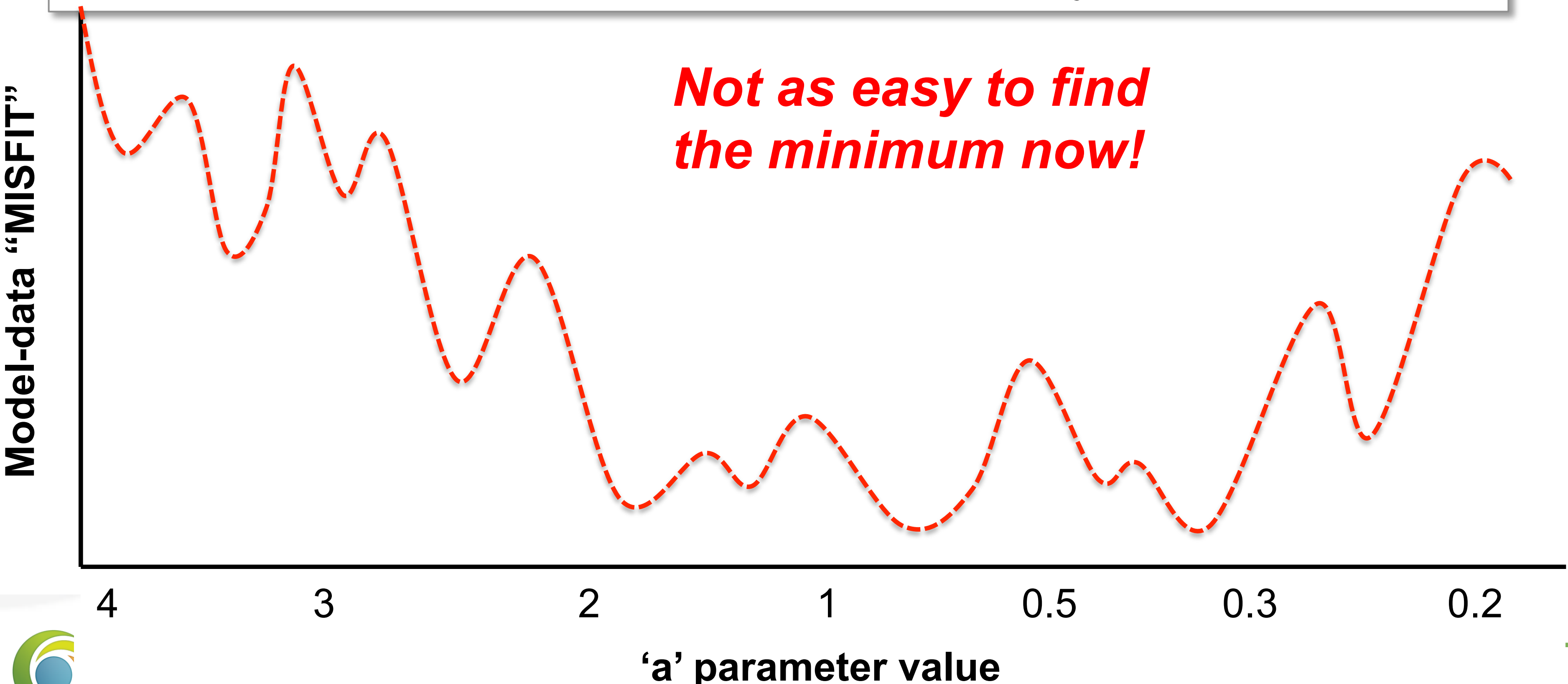
Not s

- We want … tion…
- BUT! You



Not so simple case!

- We want to find the MINIMUM of the misfit function…
- BUT! Your misfit function may look like this…!!
- How do we find the minimum numerically?



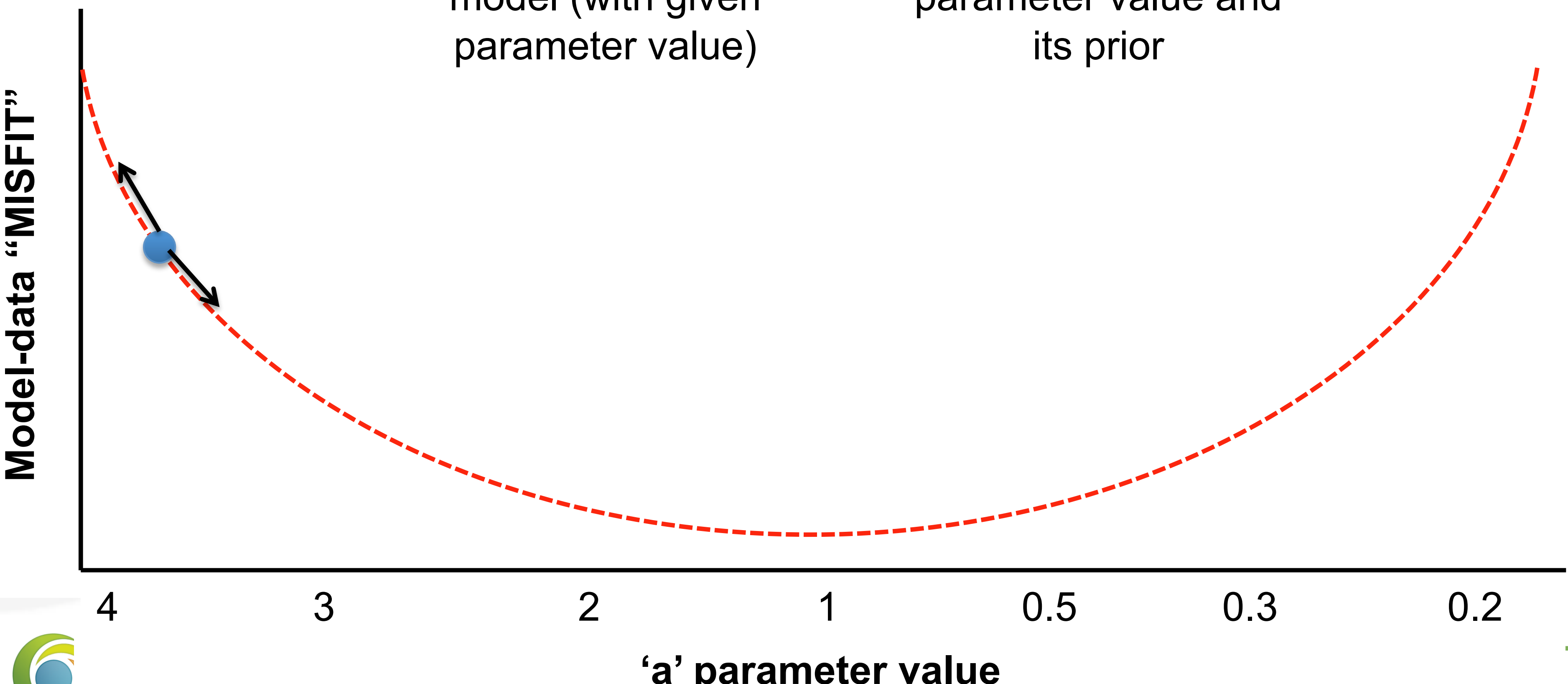
Finding the minimum…

- "Gradient-descent" methods
- Describe a "cost function":

$$J(x) = \tfrac{1}{2}(\mathbf{H}.x-y)^T\mathbf{R}^{-1}(\mathbf{H}.x-y) + \tfrac{1}{2}(x-x_b)^T\mathbf{B}^{-1}(x-x_b)$$

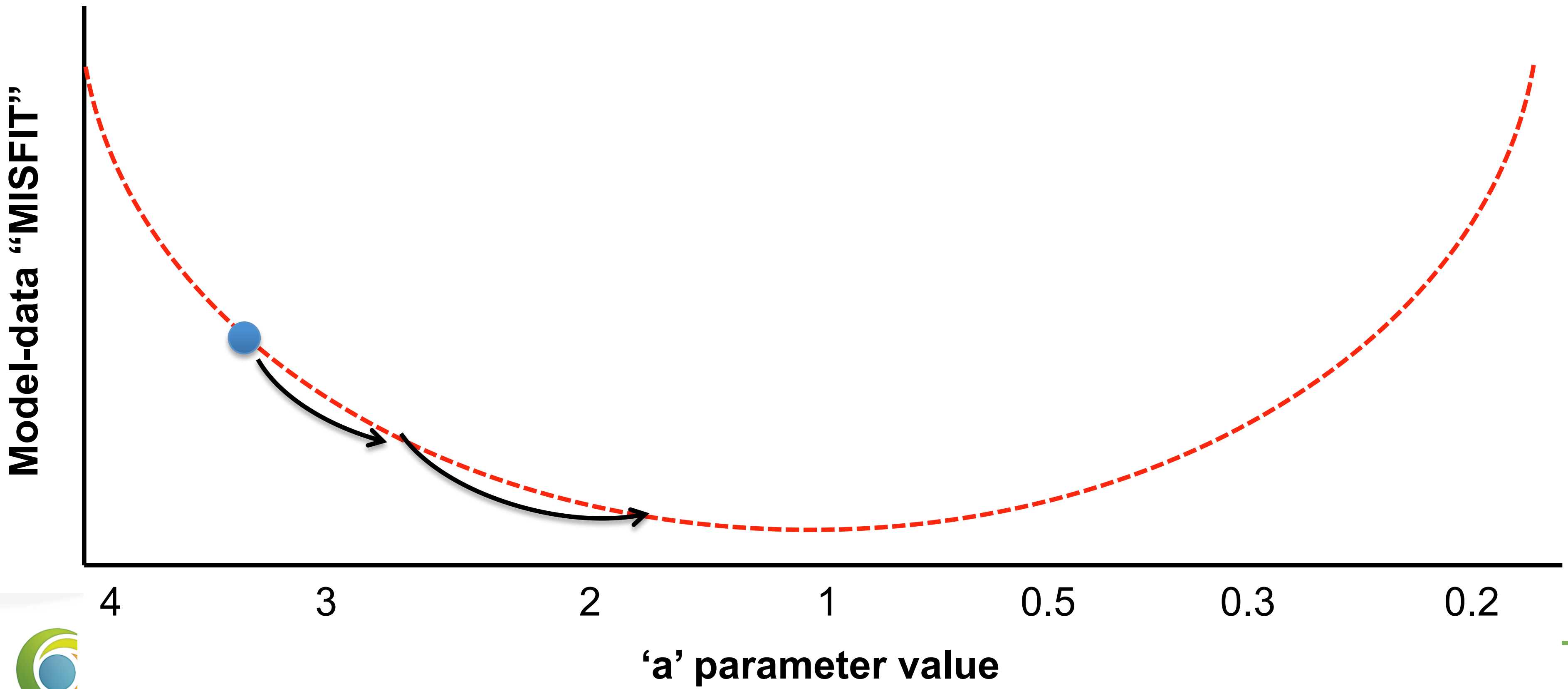
Misfit between obs. and model (with given parameter value)

Misfit between parameter value and its prior



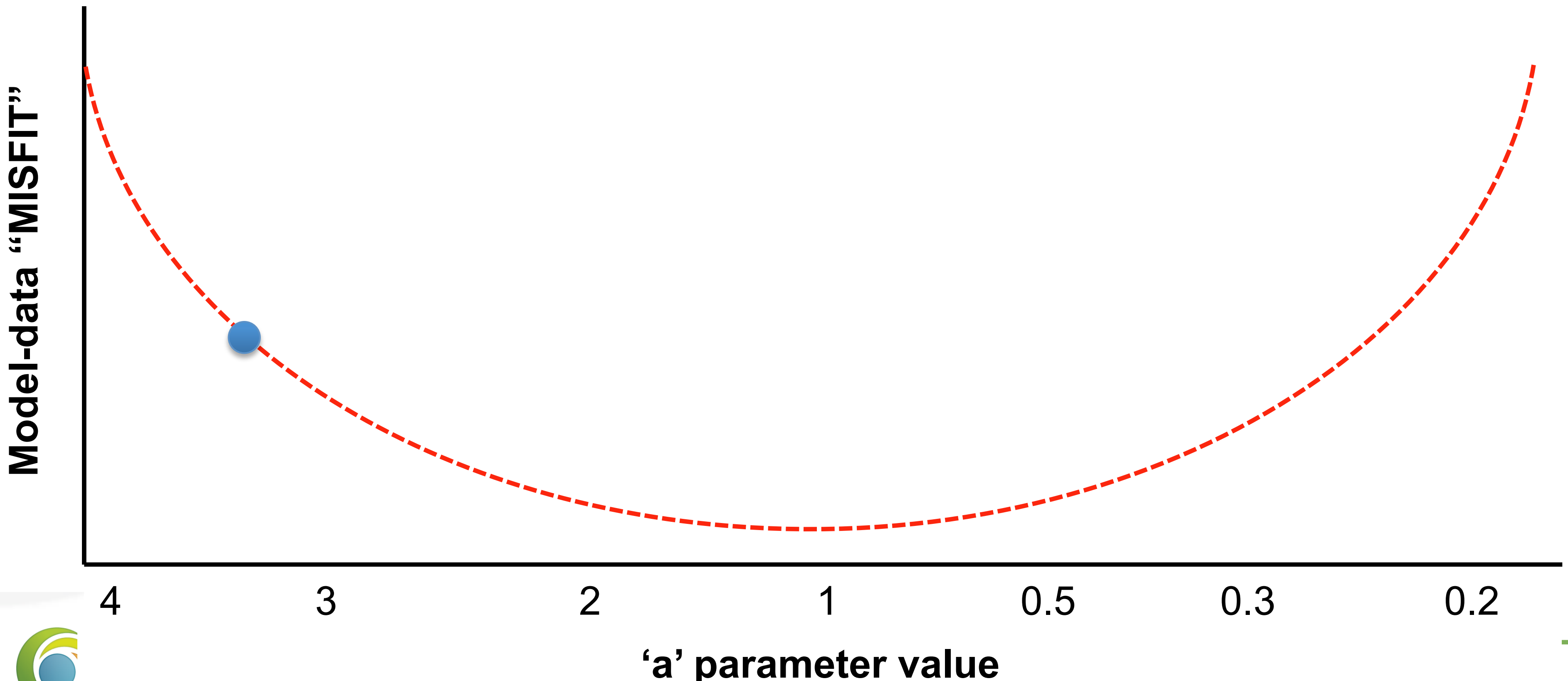
Finding the minimum…

- "Gradient-descent" methods
- Calculate the first derivative of the cost function in order to calculate the gradient…



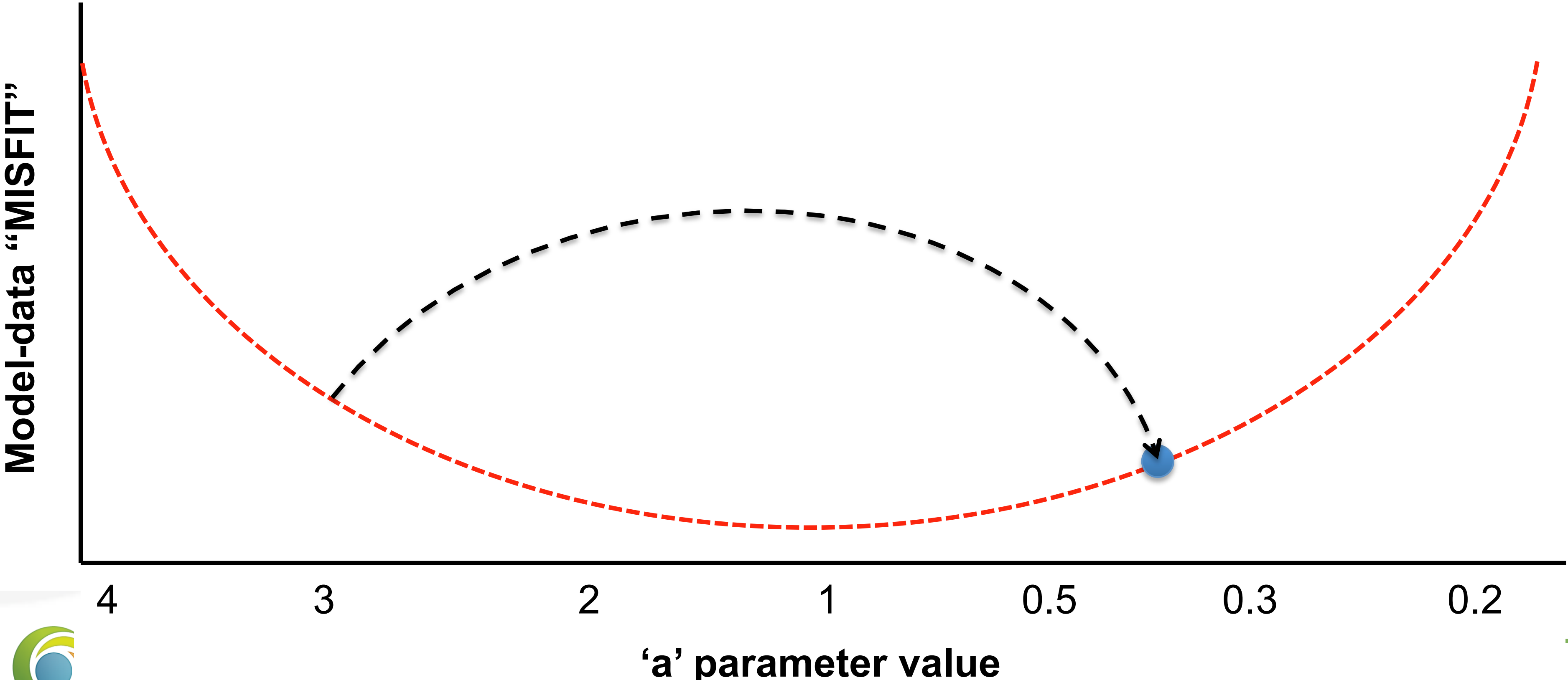
Finding the minimum…

- “Global search” methods (Genetic algorithm, Metropolis Hastings MCMC)
- Search parameter space…



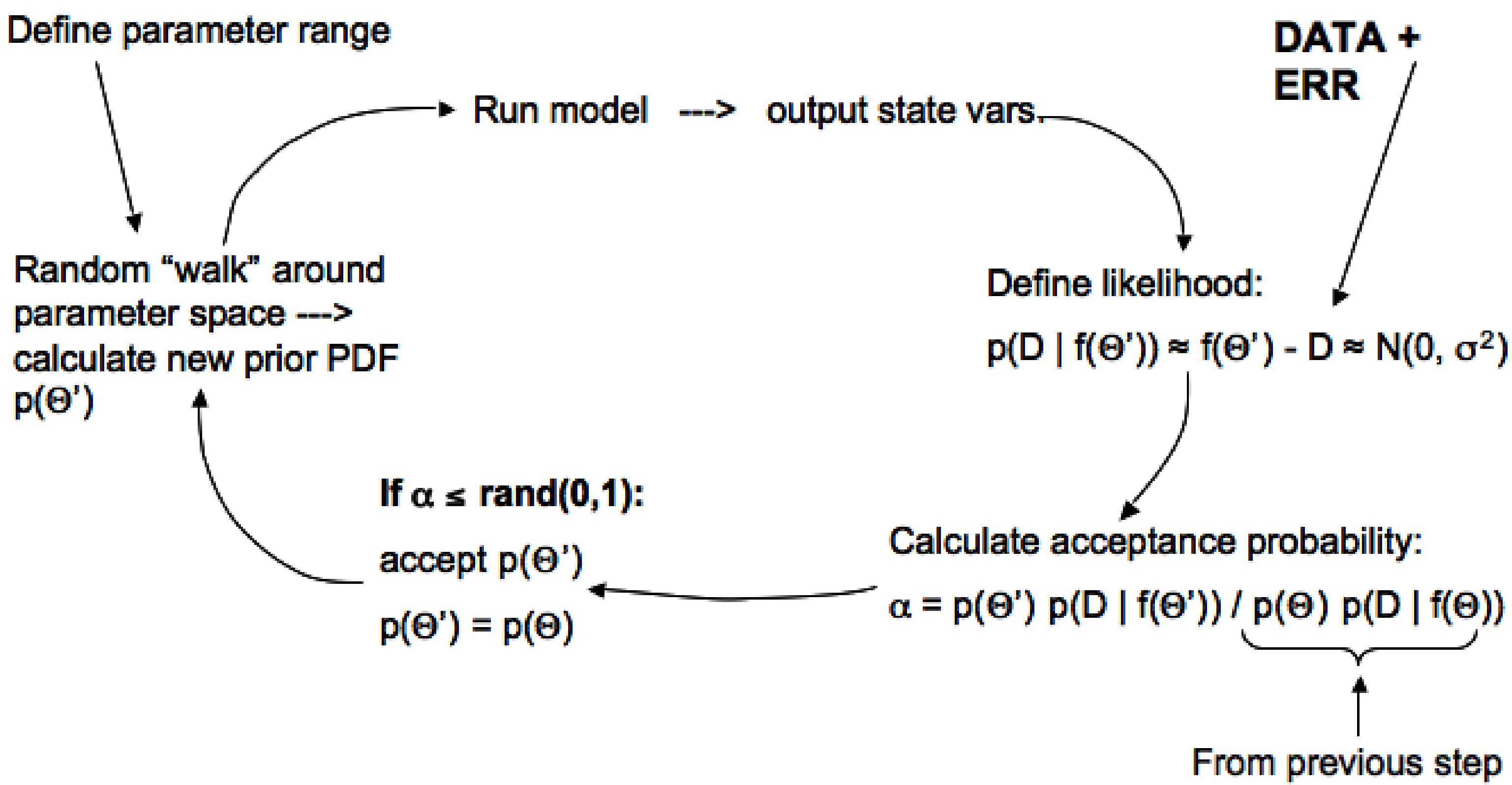
Finding the minimum…

- "Global search" methods (Genetic algorithm, Metropolis Hastings MCMC)
- Search parameter space…
- At each iteration calculate the misfit and accept or reject parameter



Finding the minimum…

- "Global search" methods (Genetic algorithm, Metropolis Hastings MCMC)
- Search parameter space…
- At each iteration calculate the misfit and accept or reject parameter

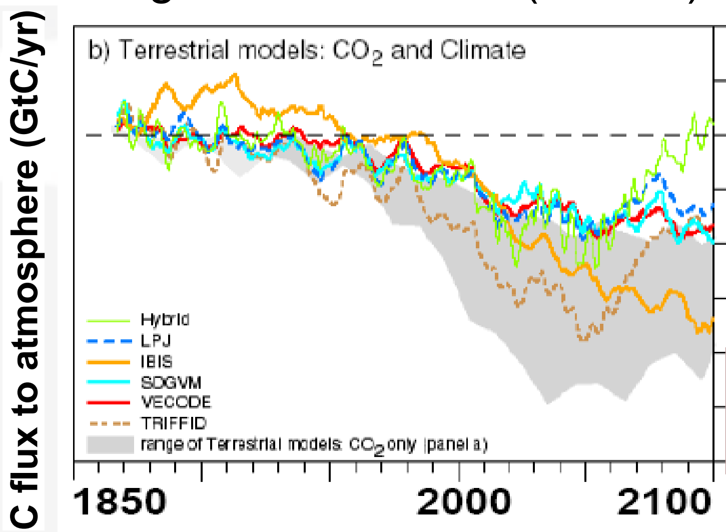


Why do we do DA?

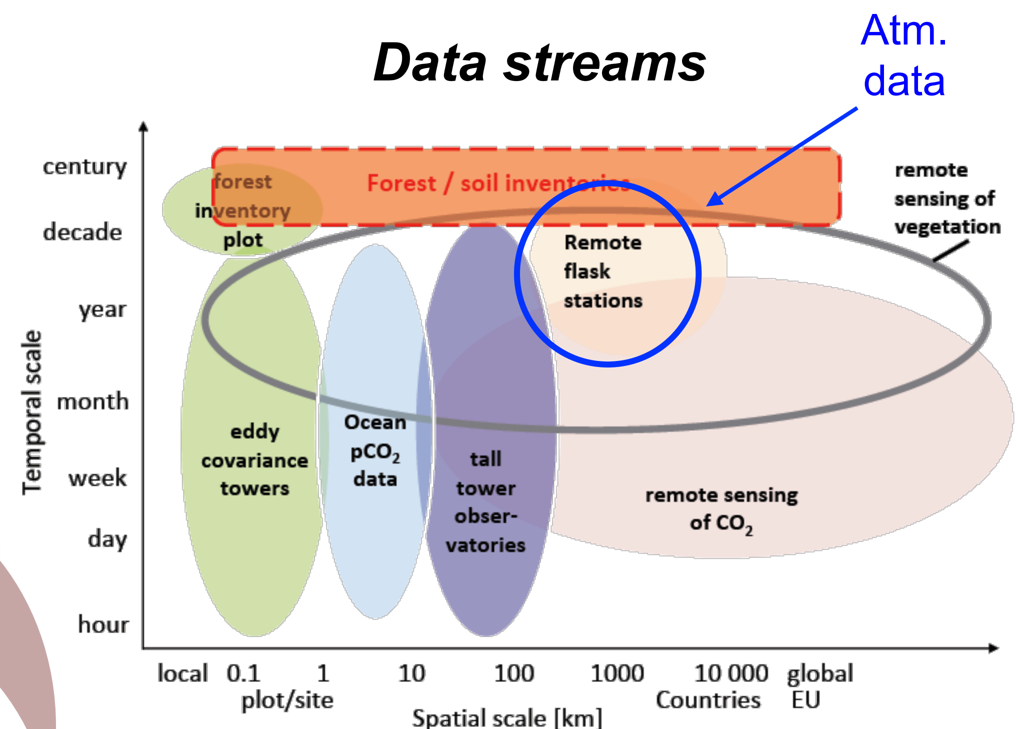
- Uncertain parameter values are one source of model error → we don't know how big a source
- Want to optimise the parameter values
- Want to get a better estimation of the uncertainty on the model simulations
  - Make predictions (C budget etc)
- *Want to improve the models → DA can help us figure out where there might be important structural errors*
- *Want to improve the DA system → other data sources, remaining issues…*

Why do we do DA?

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



Data streams



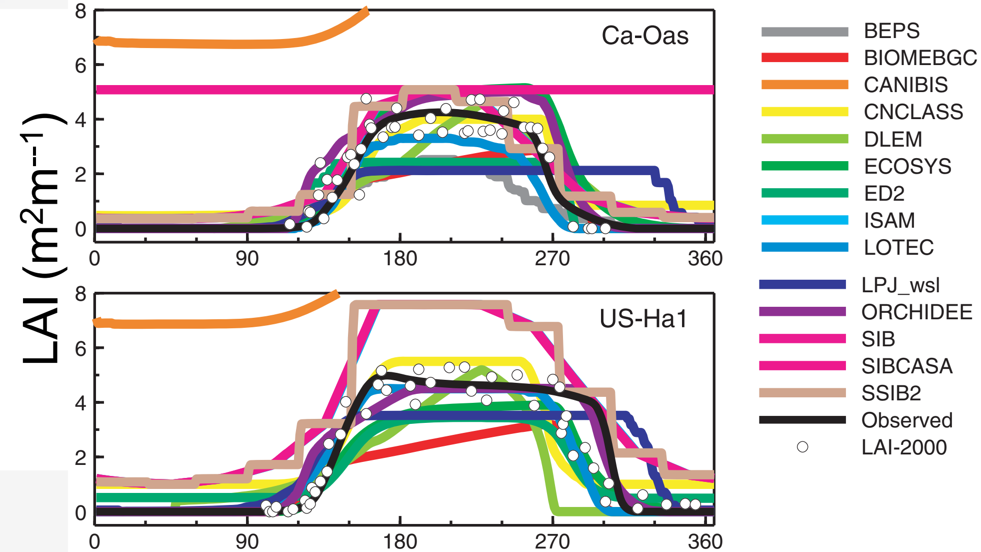
Data Assimilation

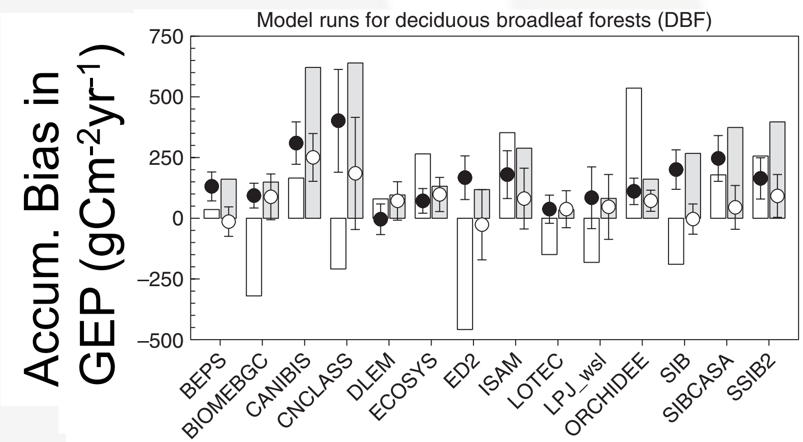
Optimized ecosystem models
➔ reduce the spread ?

Improve:

- **Process understanding**
- **Uncertainty estimates**
- **Future climate predictions**

Optimisation of the phenology





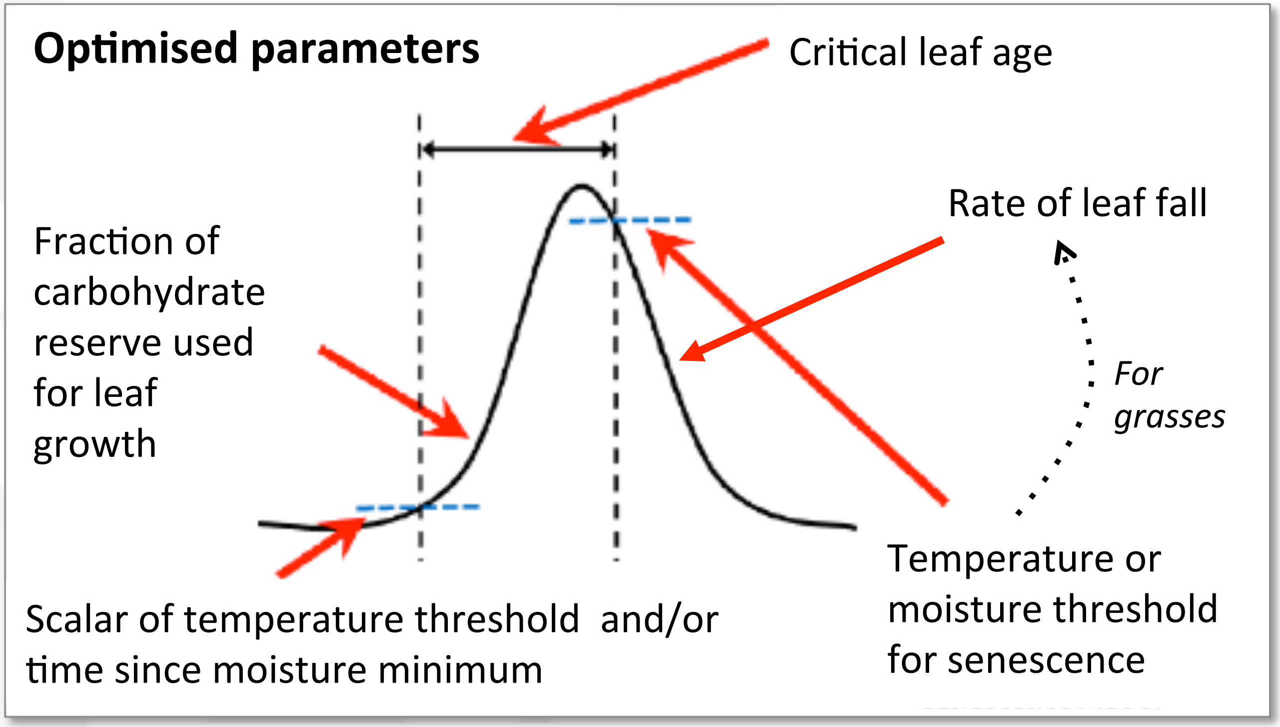
Richardson et al. (2012) GCB - NACP

- Phenology 1st order control on ecosystem fluxes
- Incorrect growing season length in TBMs in temperate/boreal regions
- Can this be improved with parameter optimisation?
- Poorer understanding and representation of leaf phenology in tropics

Leaf phenology in ORCHIDEE

	ONSET	SENESCENCE Critical leaf age +
Temperate/Boreal Deciduous	Temperature-related threshold (GDD+NCD, NGD)	Temperature threshold
Tropical raingreen	Moisture-related threshold (time since moisture minimum)	Moisture threshold
C3 and C4 grasses	Temperature-related threshold (GDD) + Moisture-related threshold	Temperature threshold + Moisture threshold

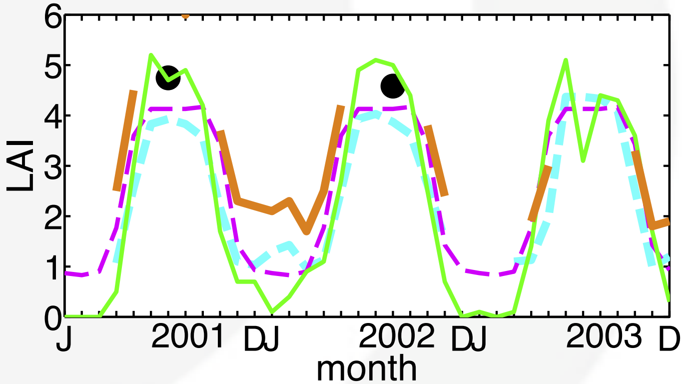
Carbon Cycle Data Assimilation System (CCDAS)



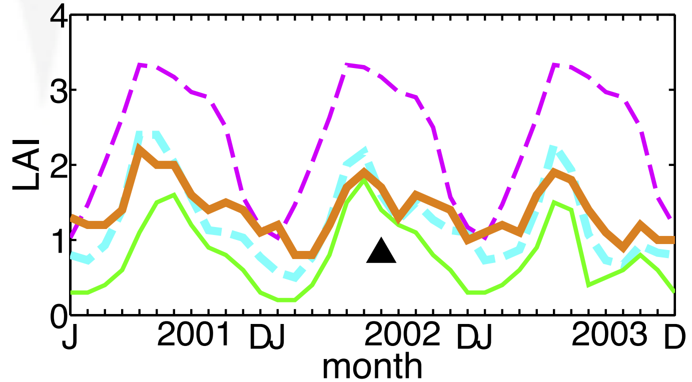
- 4 – 6 parameters per PFT
- 15 random grid points with available obs.
- PFT vegetation cover > 0.6
- Multi-site and single-site optim.
- 4D variational + finite difference approach

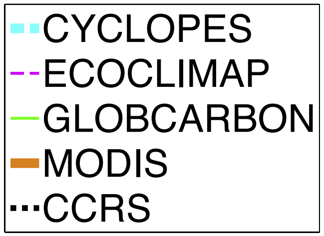
NDVI from satellite reflectance data

- MODIS collection 5 5km surface reflectance data (2000-2008)
- Corrected for directional effects (Vermote et al., 2009)
- Averaged at model grid scale, interpolated to daily timeseries
- Model LAI to fAPAR using simple Beer-Lambert Law
- Normalise MODIS NDVI / modelled fAPAR → assumption of linear relationship (5 – 95th percentile)



Deciduous broadleaf forest



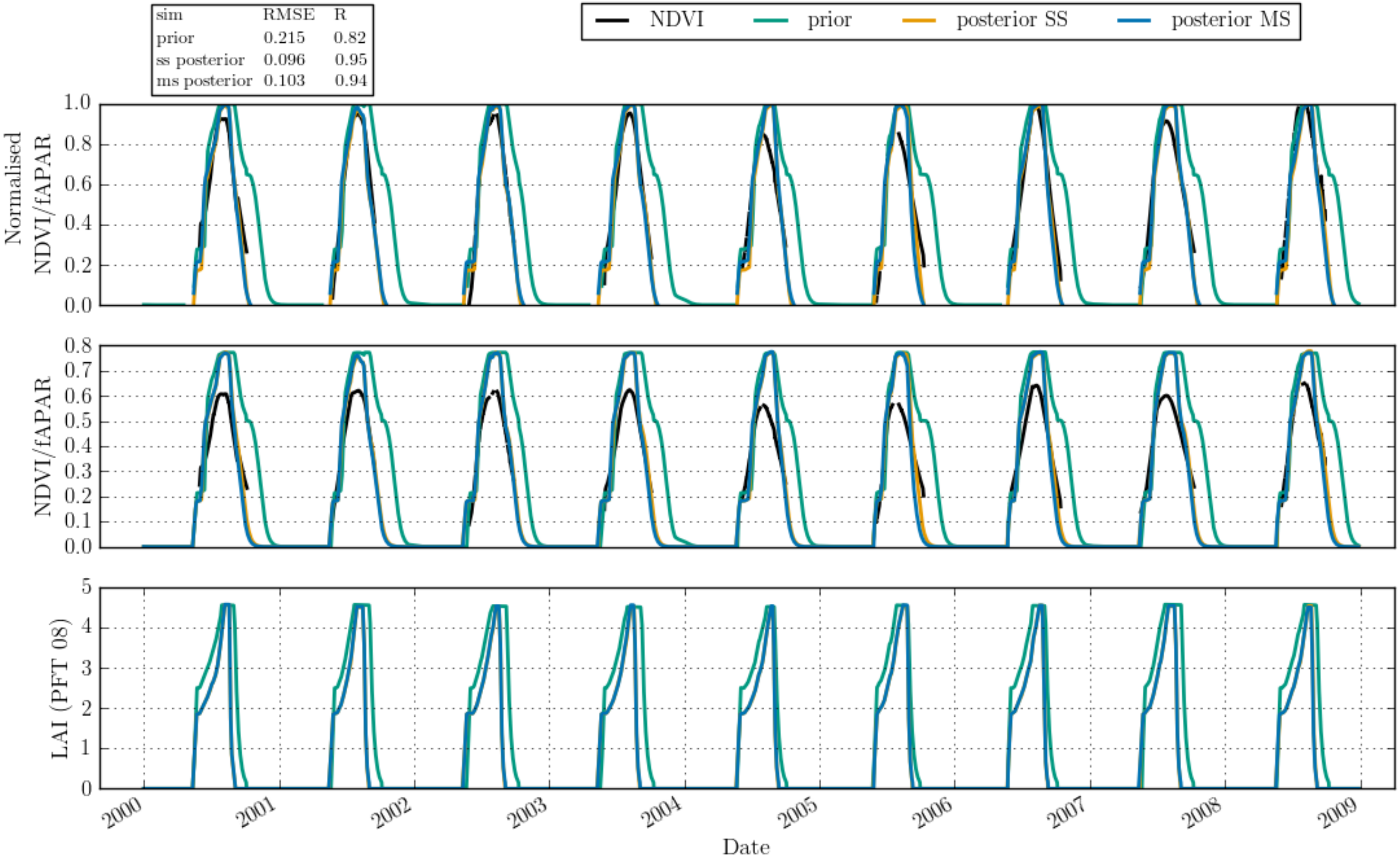


Grassland

Garrigues et al. (2008) JGR

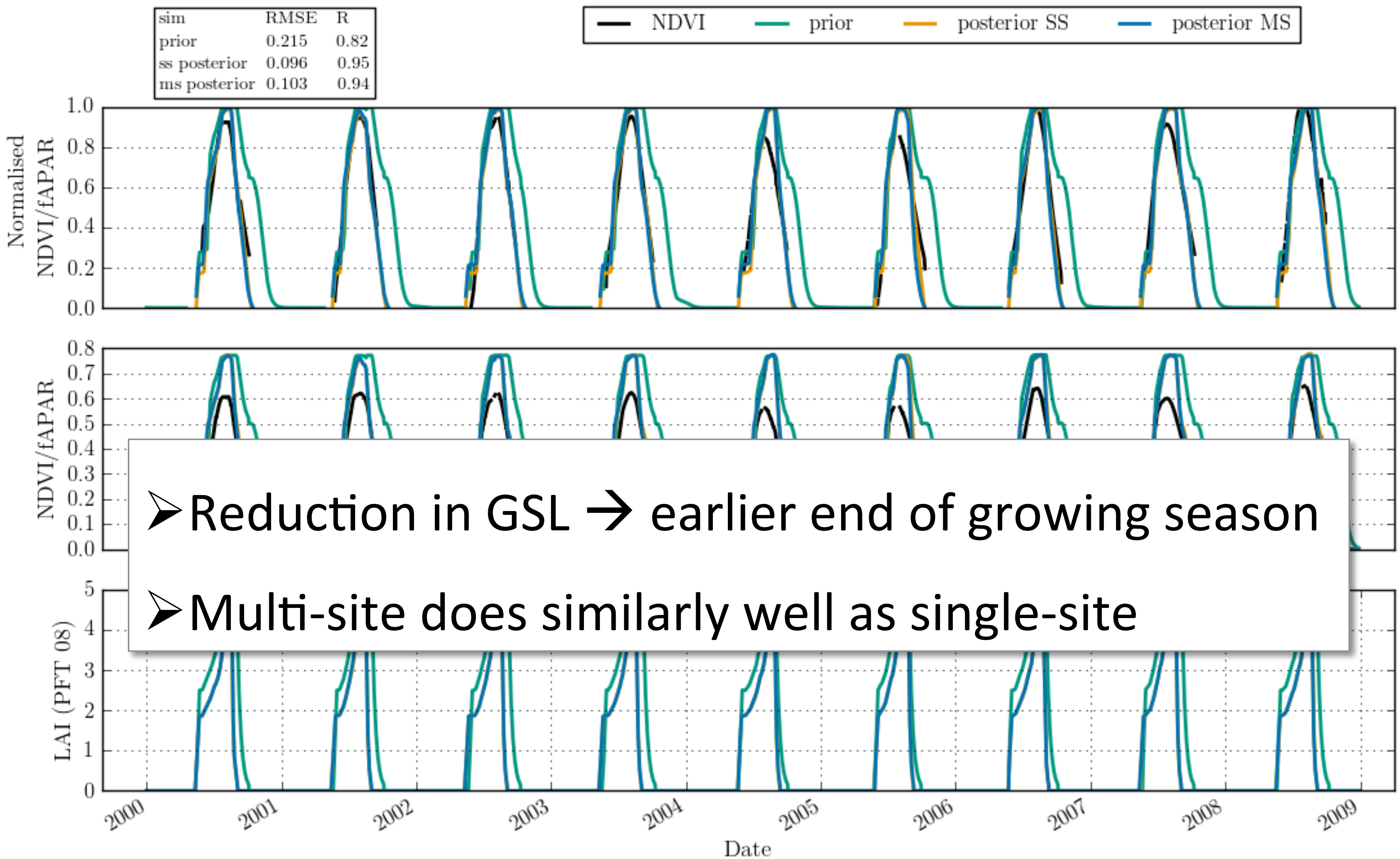
Temperate and boreal deciduous forest

Example: Boreal broadleaved deciduous



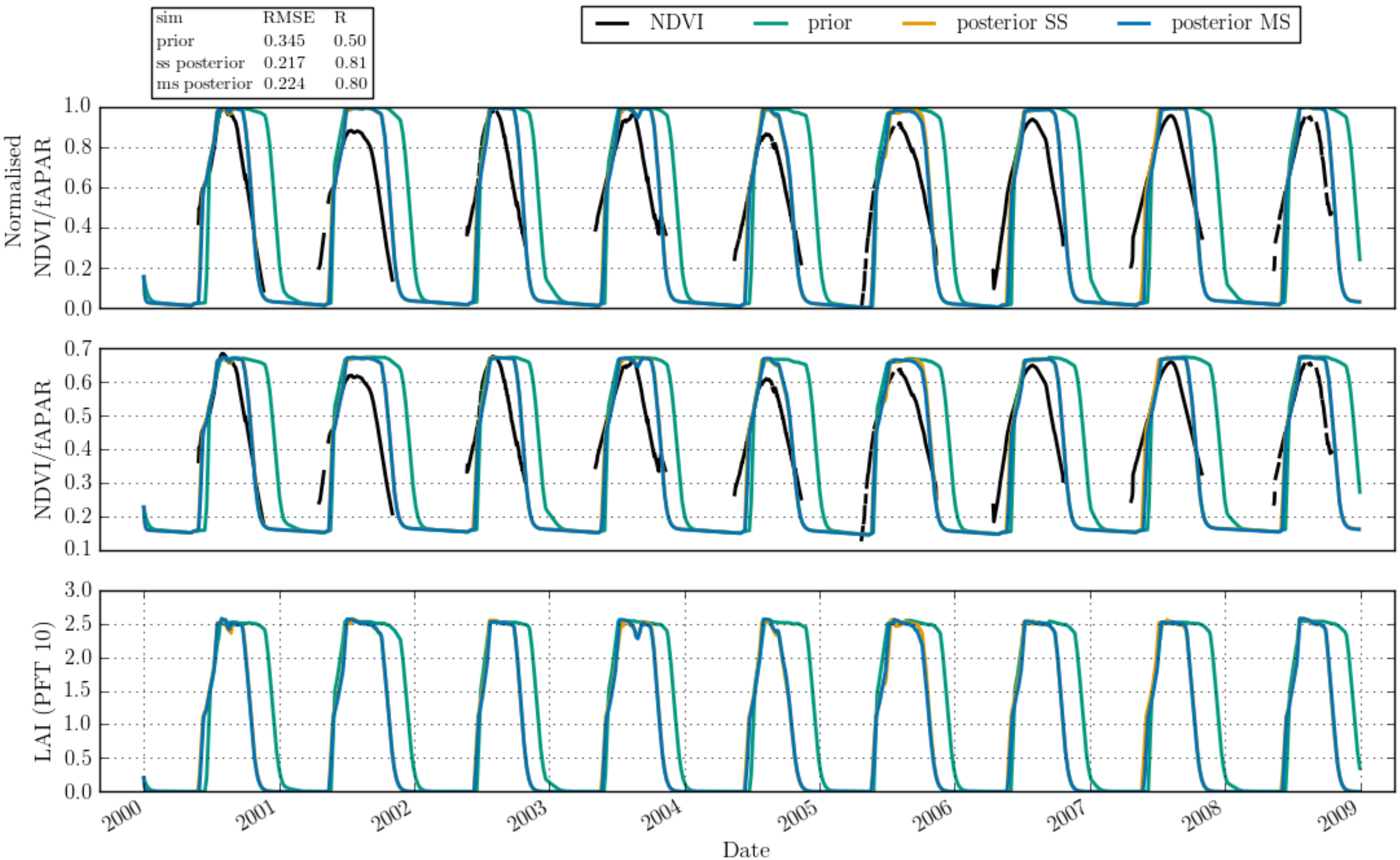
Temperate and boreal deciduous forest

Example: Boreal broadleaved deciduous

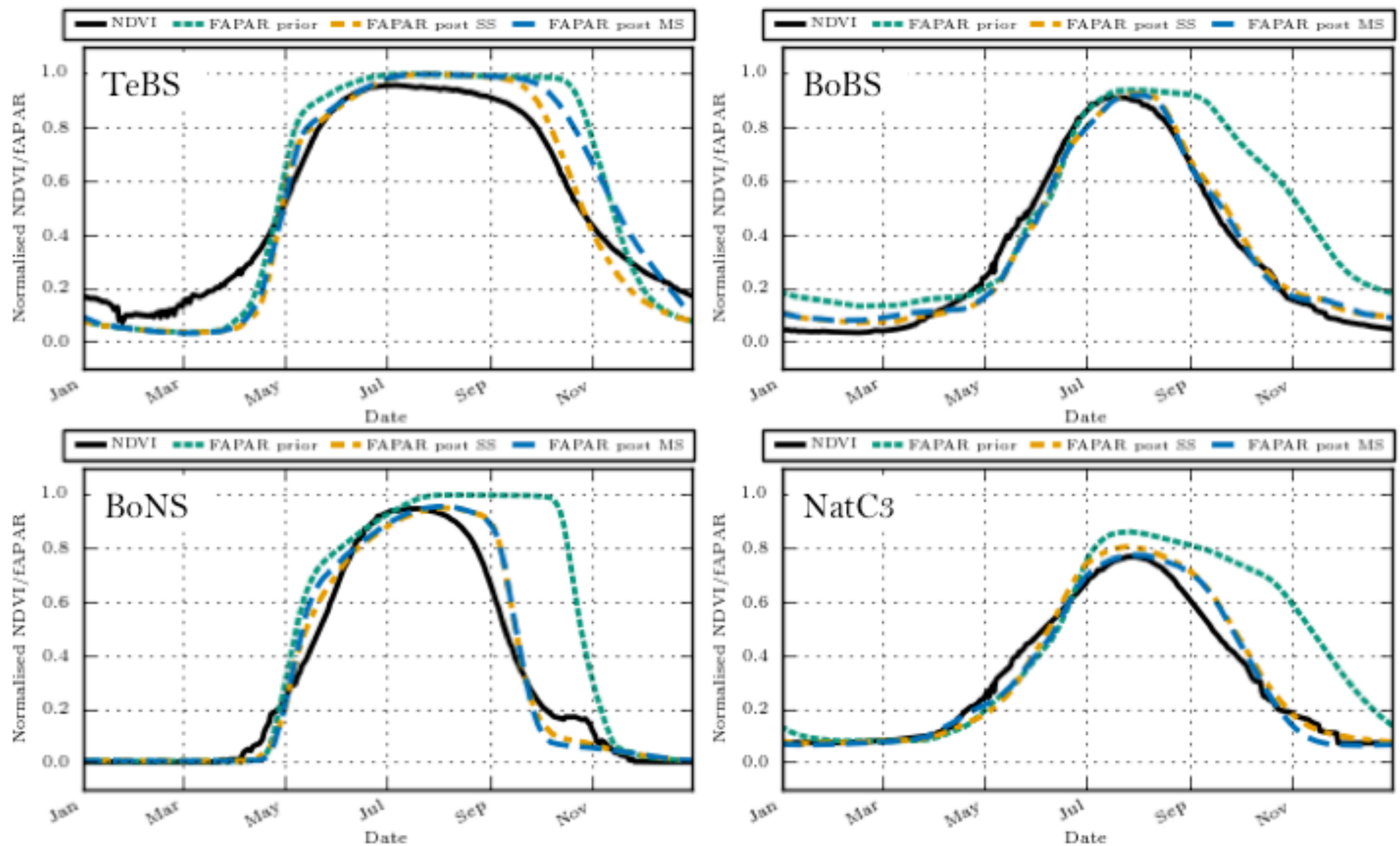


- ➢Reduction in GSL → earlier end of growing season
- ➢Multi-site does similarly well as single-site

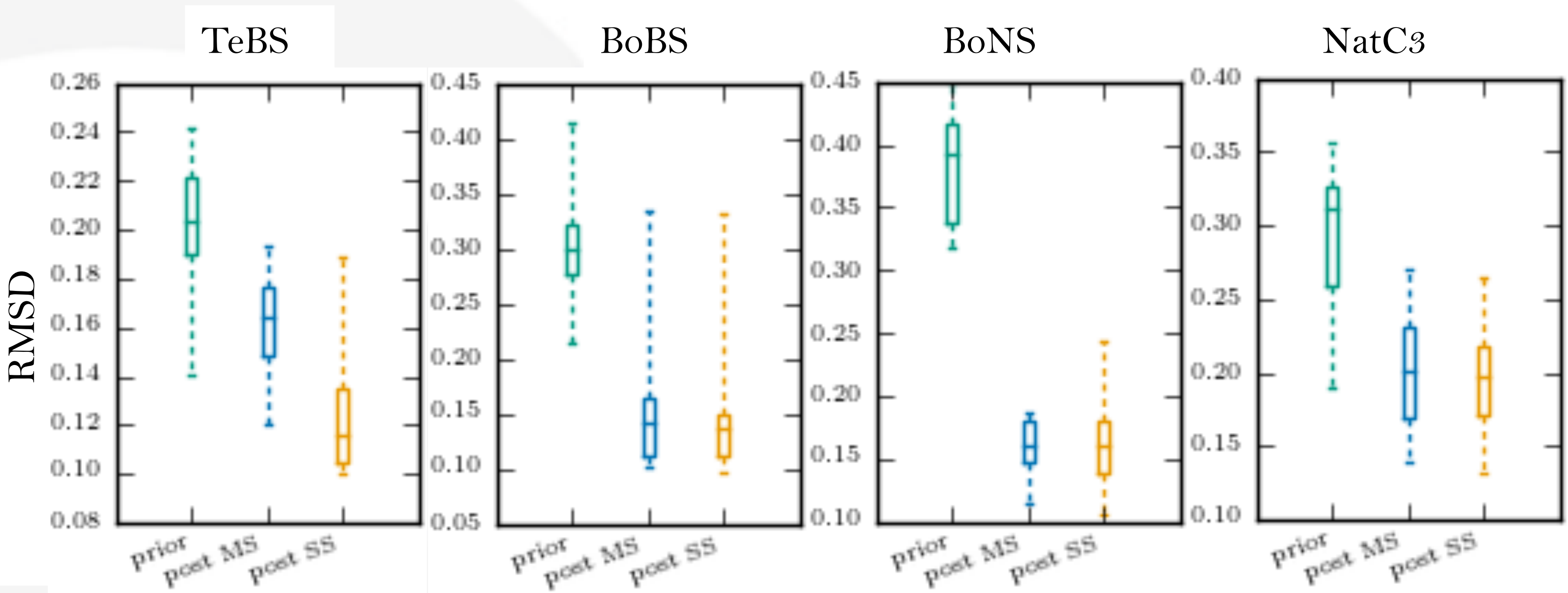
Natural C3 grass

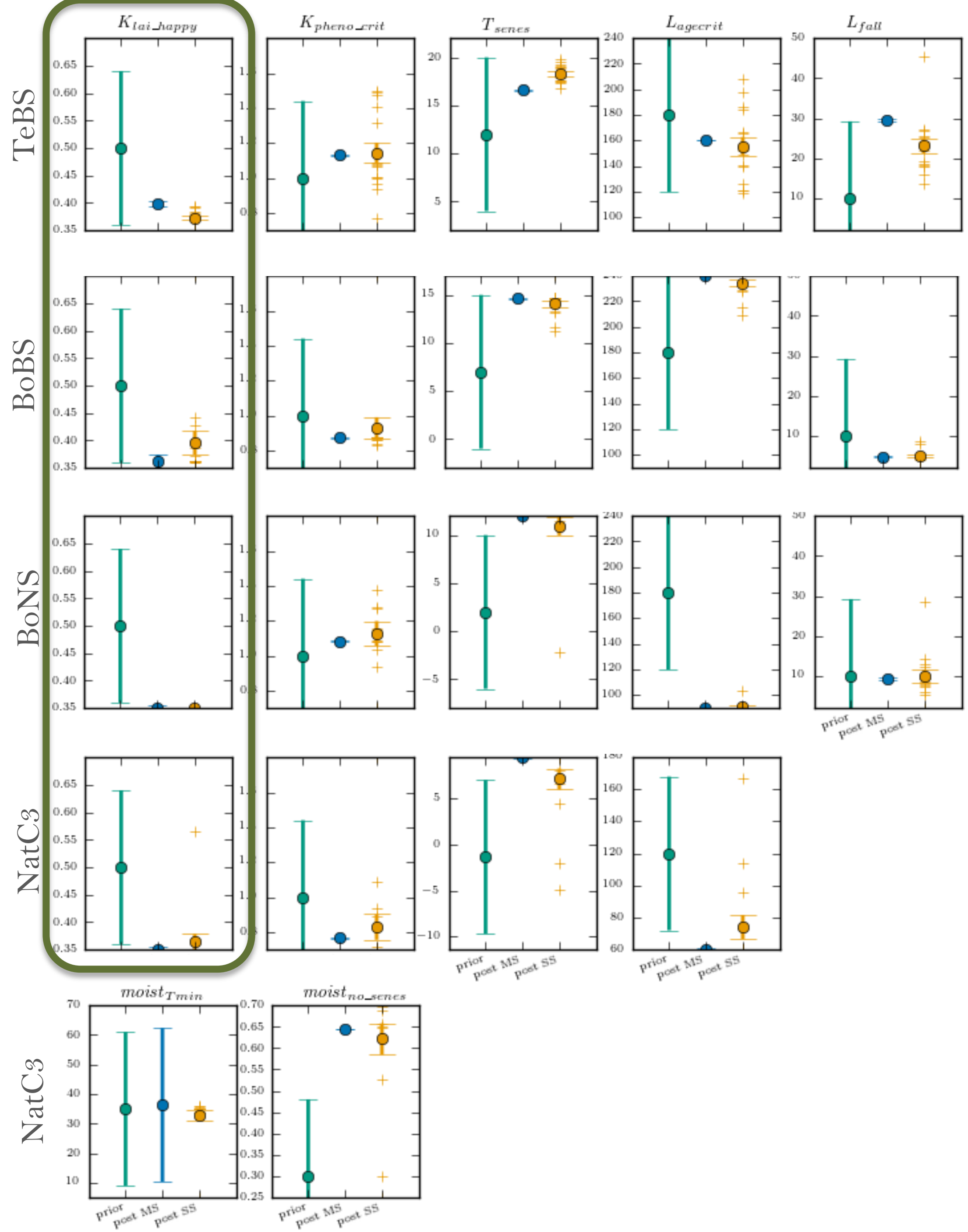


Mean seasonal cycle



Mean seasonal cycle

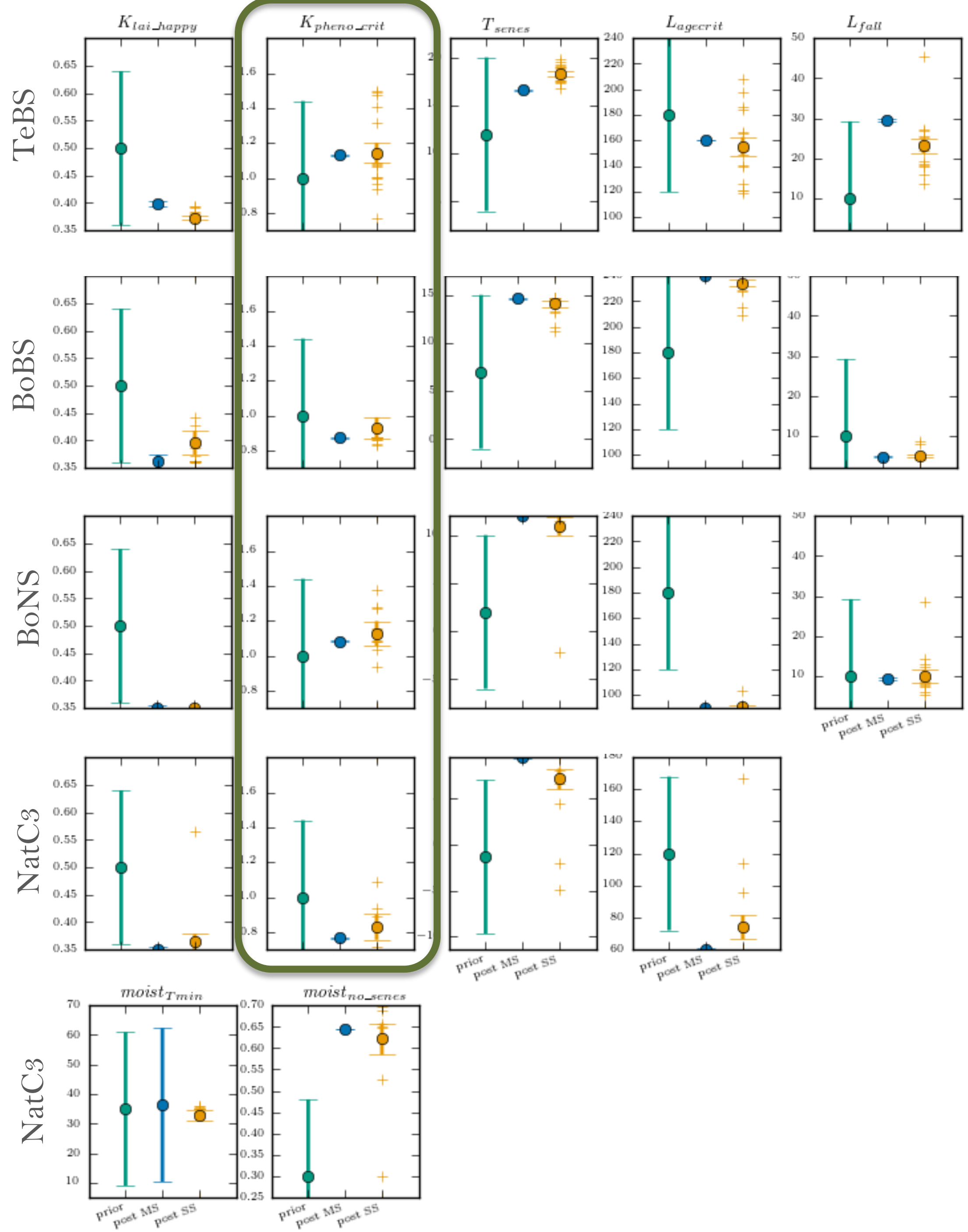




Posterior parameters

ONSET

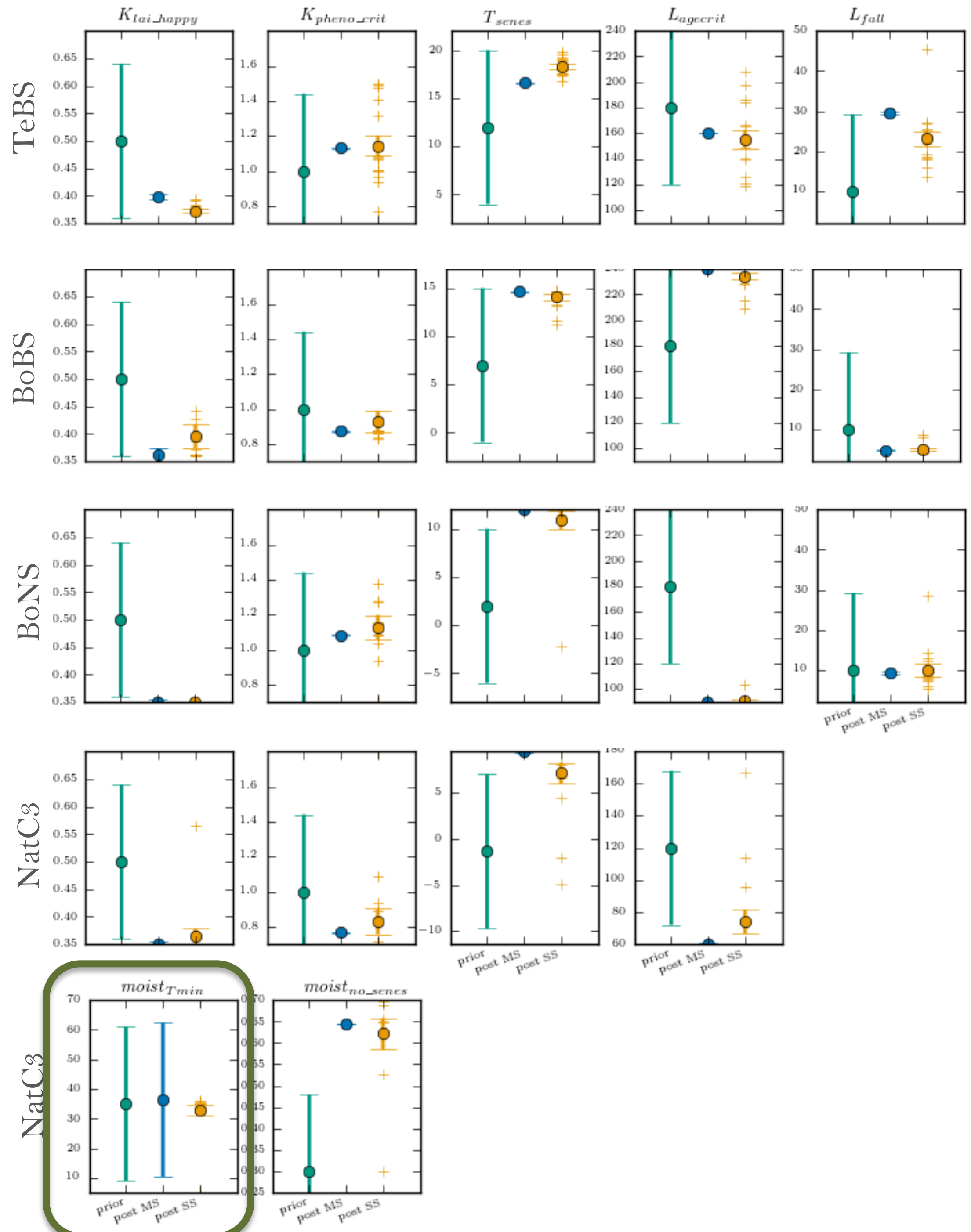
- Decrease in fraction of carbohydrate reserve for leaf growth



Posterior parameters

ONSET

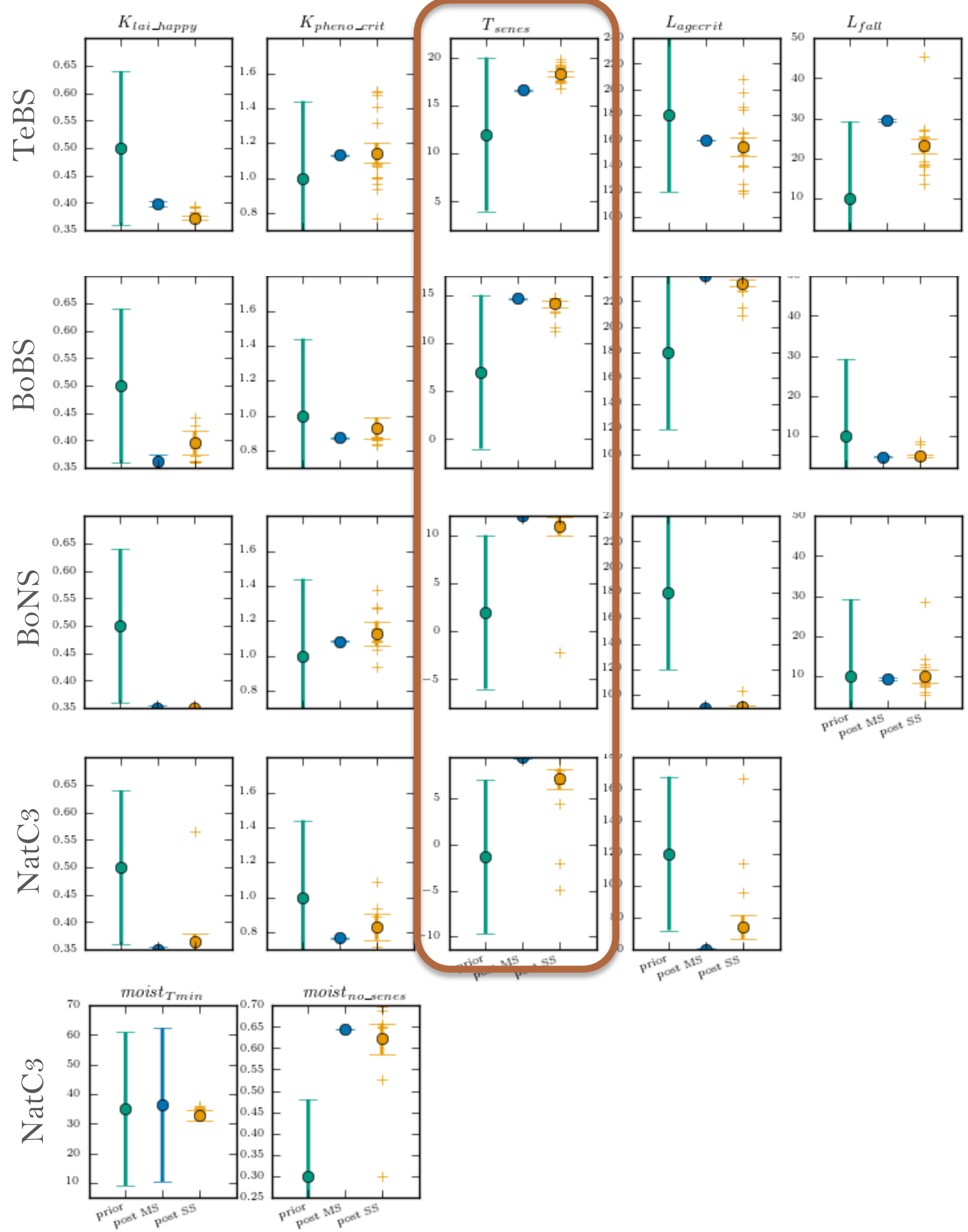
- Decrease in fraction of carbohydrate reserve for leaf growth
- Scalar on GDD / NGD leaf onset threshold – not strong constraint



Posterior parameters

ONSET

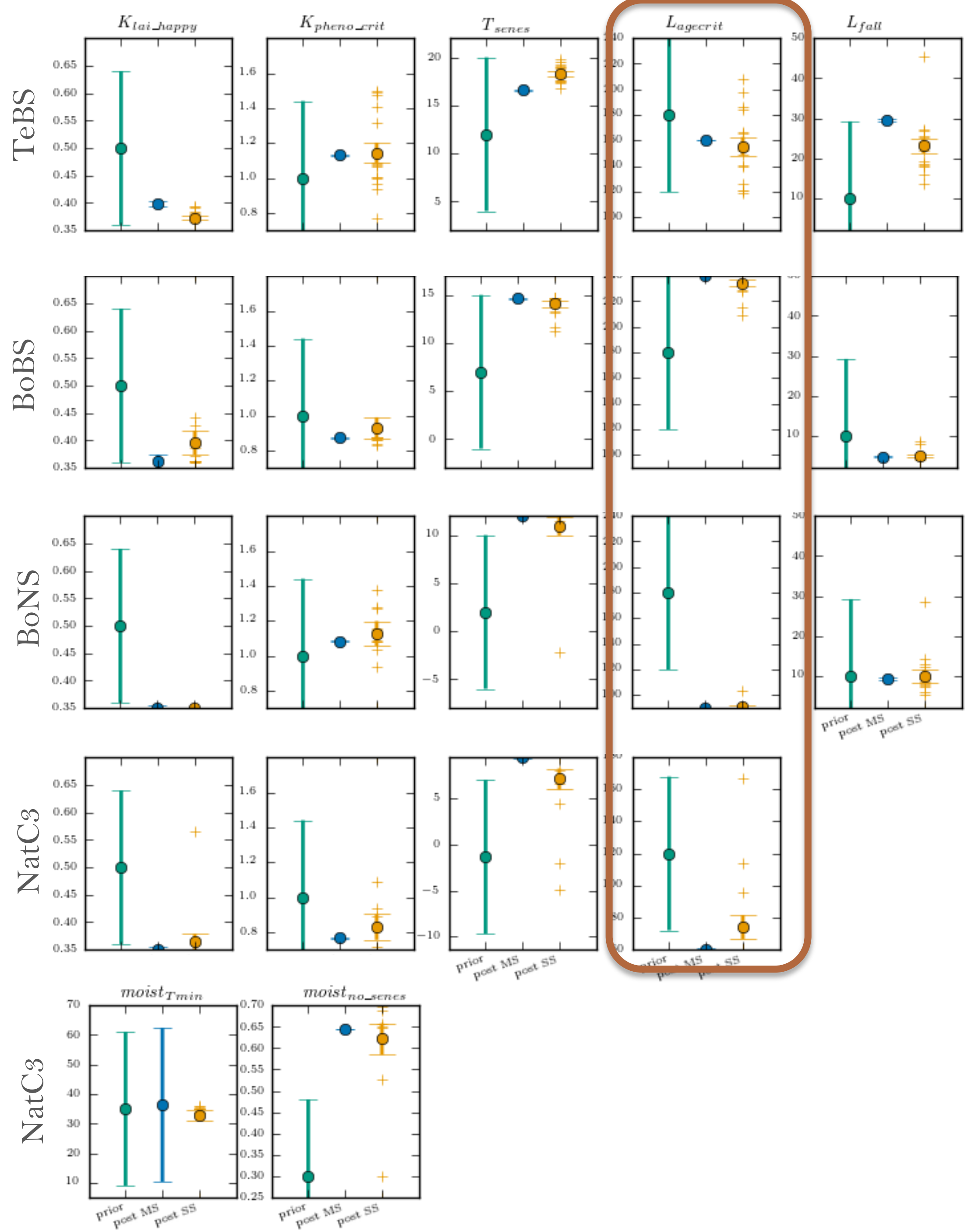
- Decrease in fraction of carbohydrate reserve for leaf growth
- Scalar on GDD / NGD leaf onset threshold – not strong constraint
- Minimum time since moisture minimum – no constraint for C3 grasses



Posterior parameters

SENESCENCE

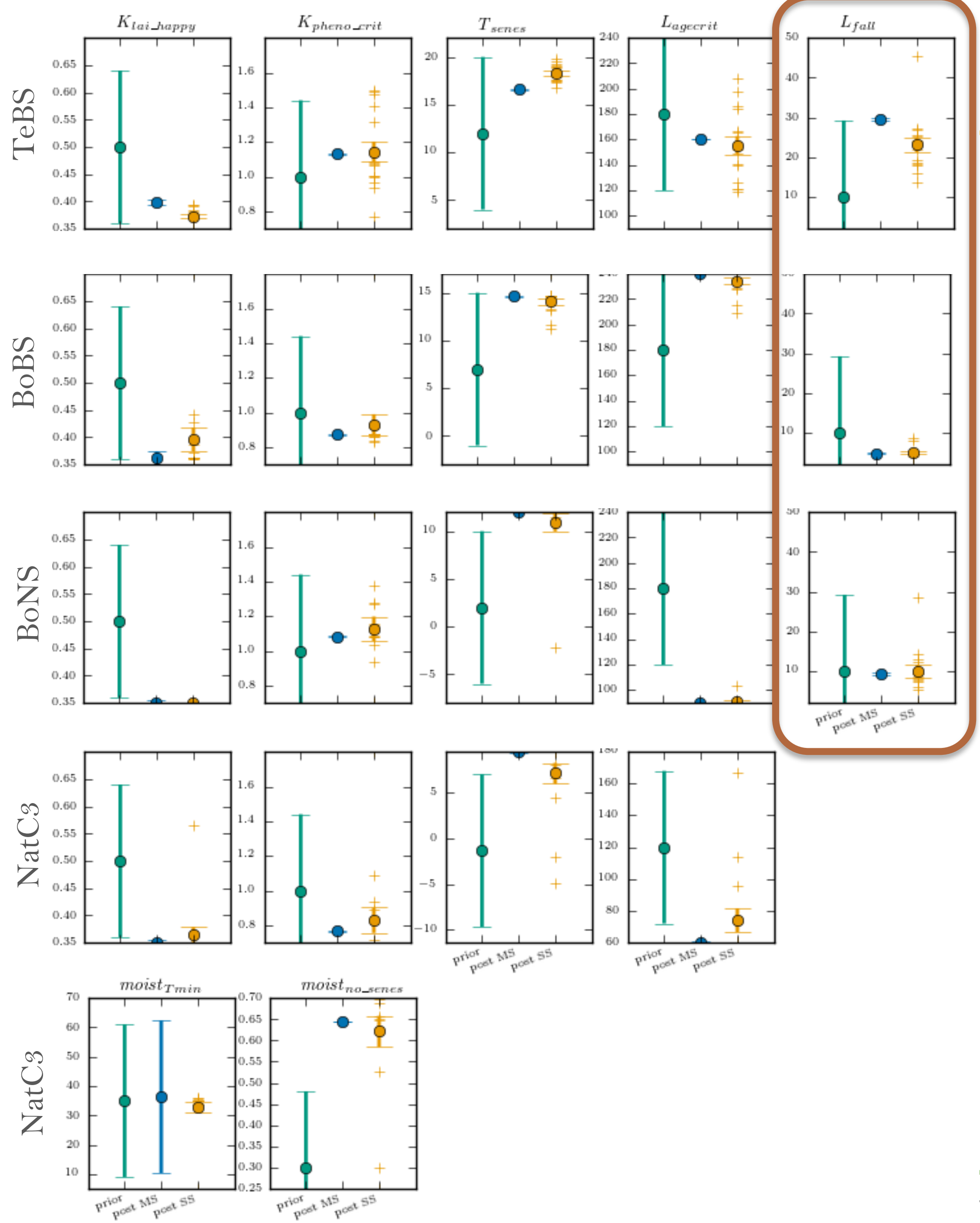
- Increase in temperature threshold for senescence



Posterior parameters

SENESCENCE

- Increase in temperature threshold for senescence
- Some decrease in critical leaf age for senescence



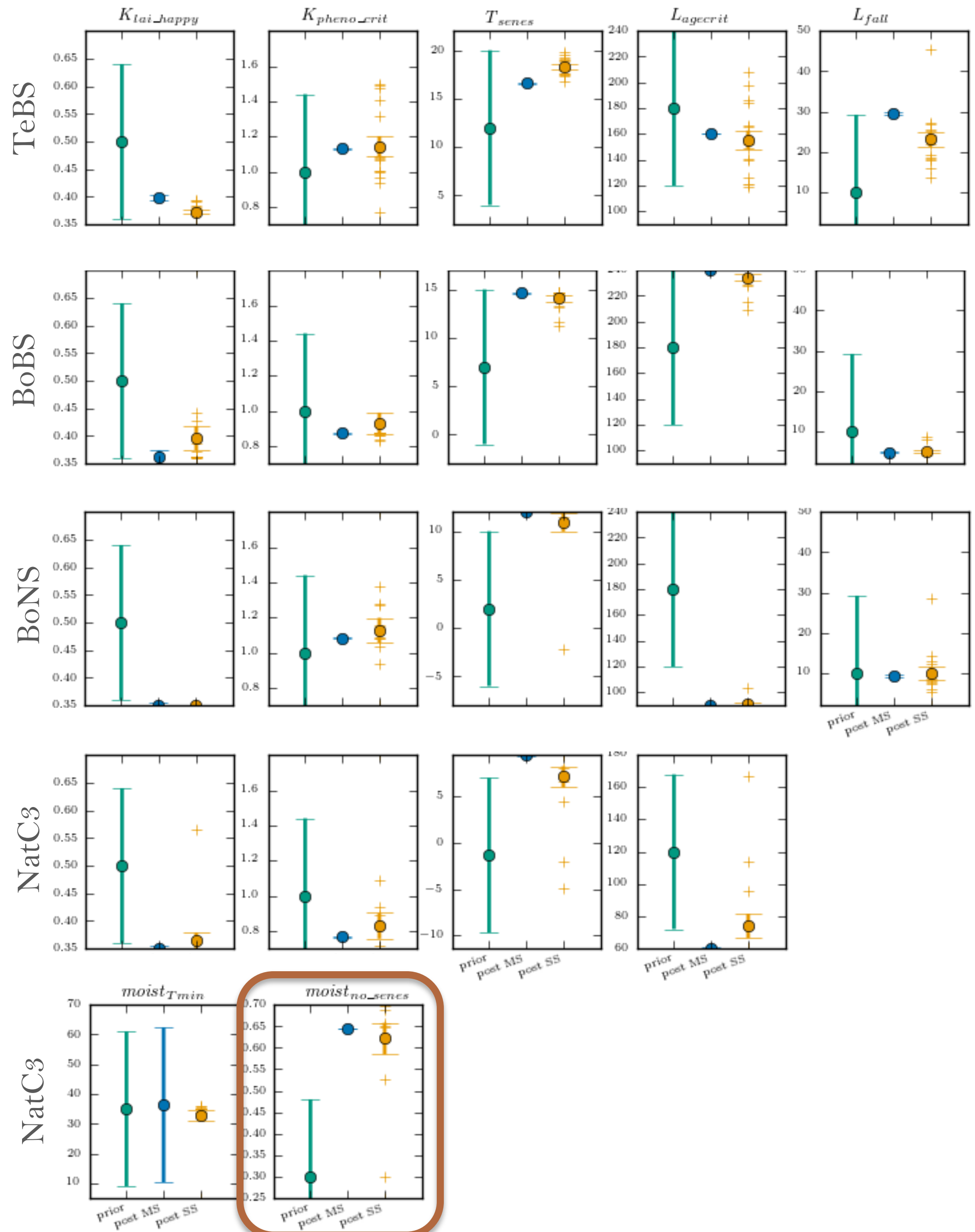
Posterior parameters

SENESCENCE

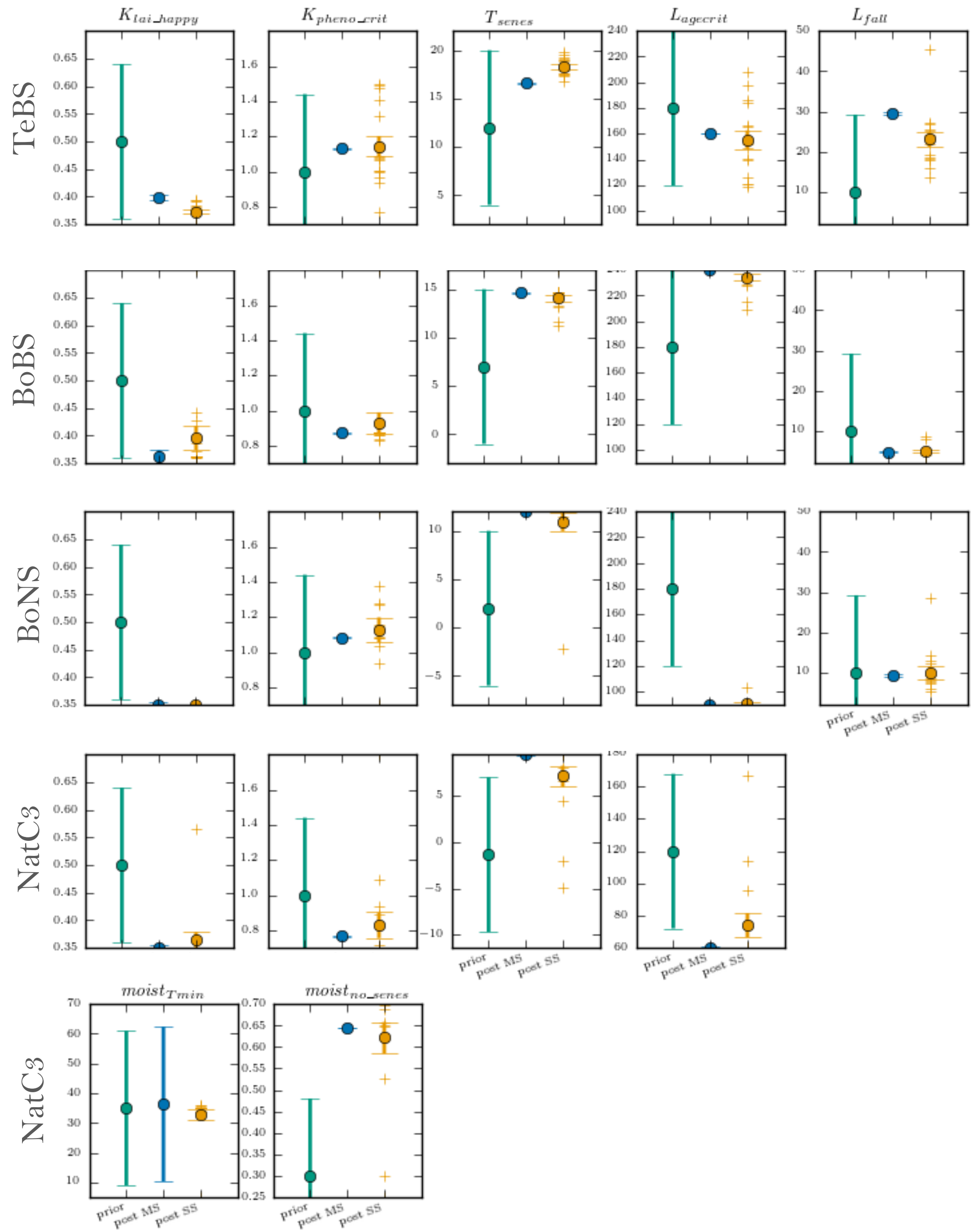
- Increase in temperature threshold for senescence
- Some decrease in critical leaf age for senescence
- Some decrease in the rate of leaf fall in autumn

Posterior parameters

SENESCENCE



- Increase in temperature threshold for senescence
- Some decrease in critical leaf age for senescence
- Some decrease in the rate of leaf fall in autumn
- Increase in the moisture threshold for senescence for C3 grasses

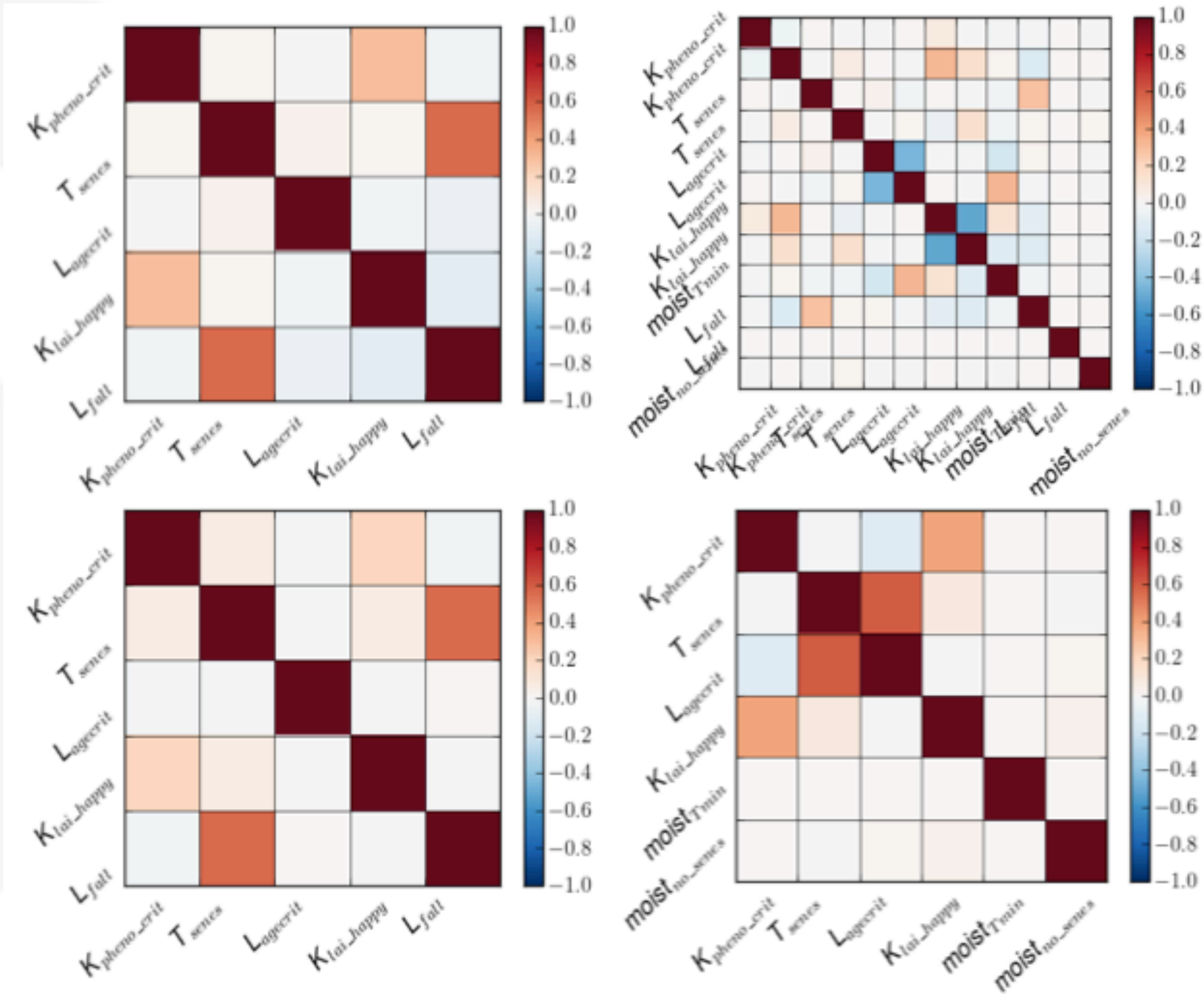


Posterior parameters

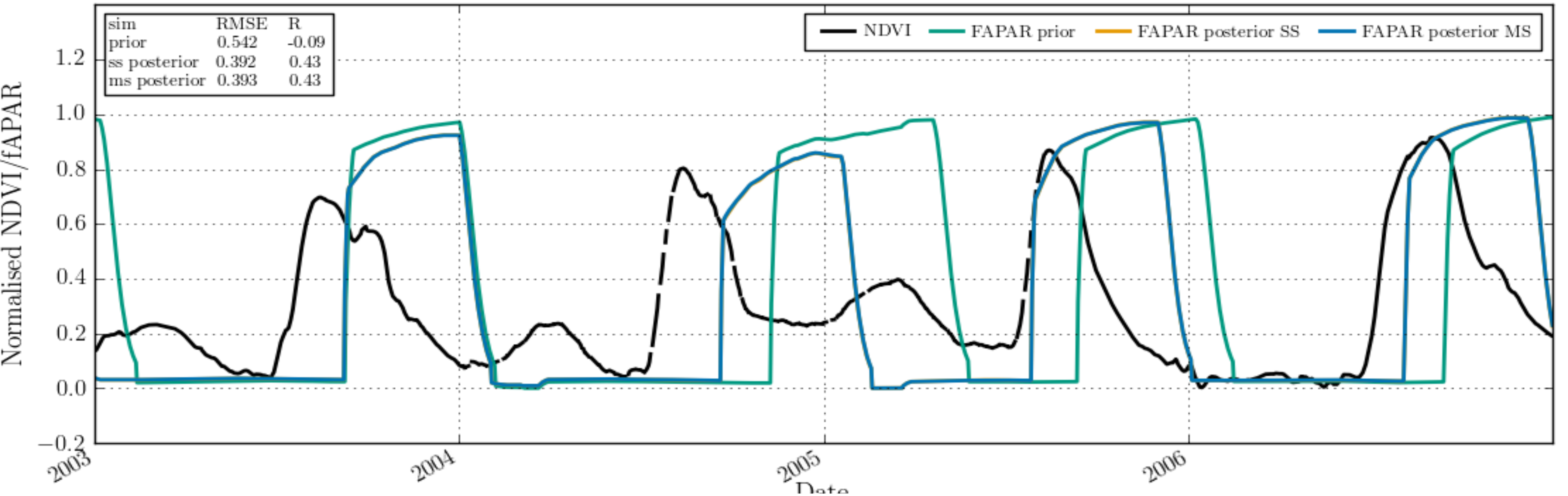
Discussion points…

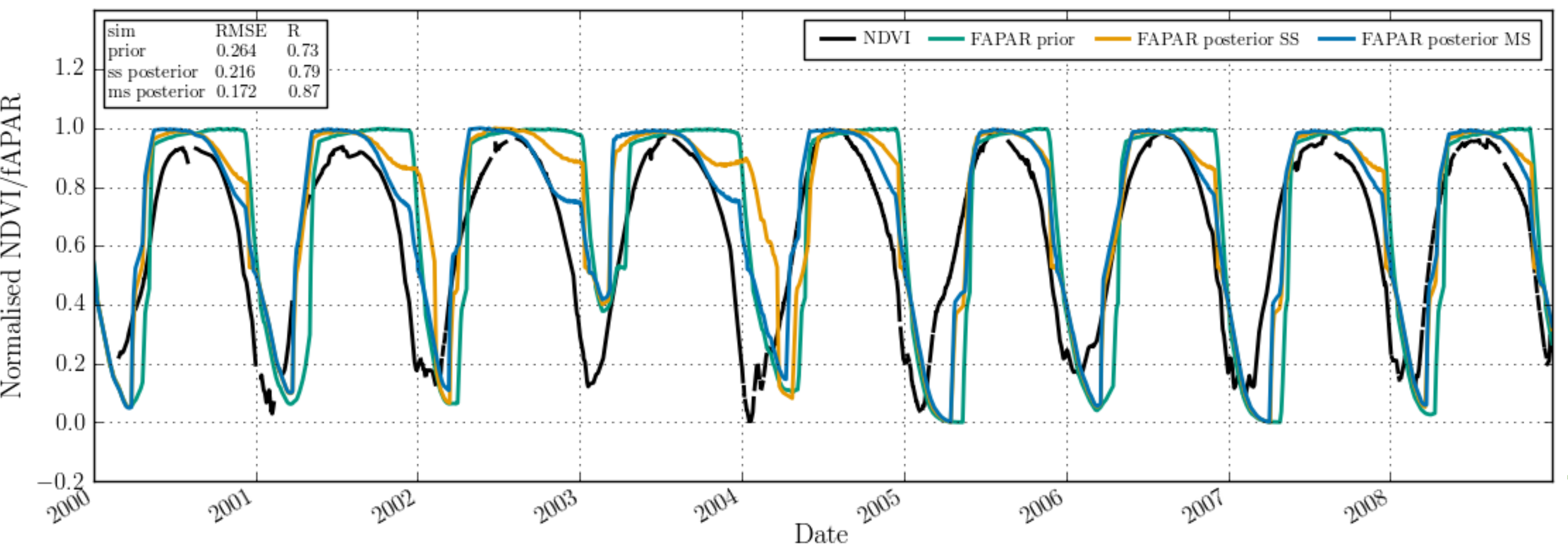
- Are these values realistic?
- Do we really get an idea of which processes are missing?
- Edge-hitting parameters?

Posterior parameters – covariance

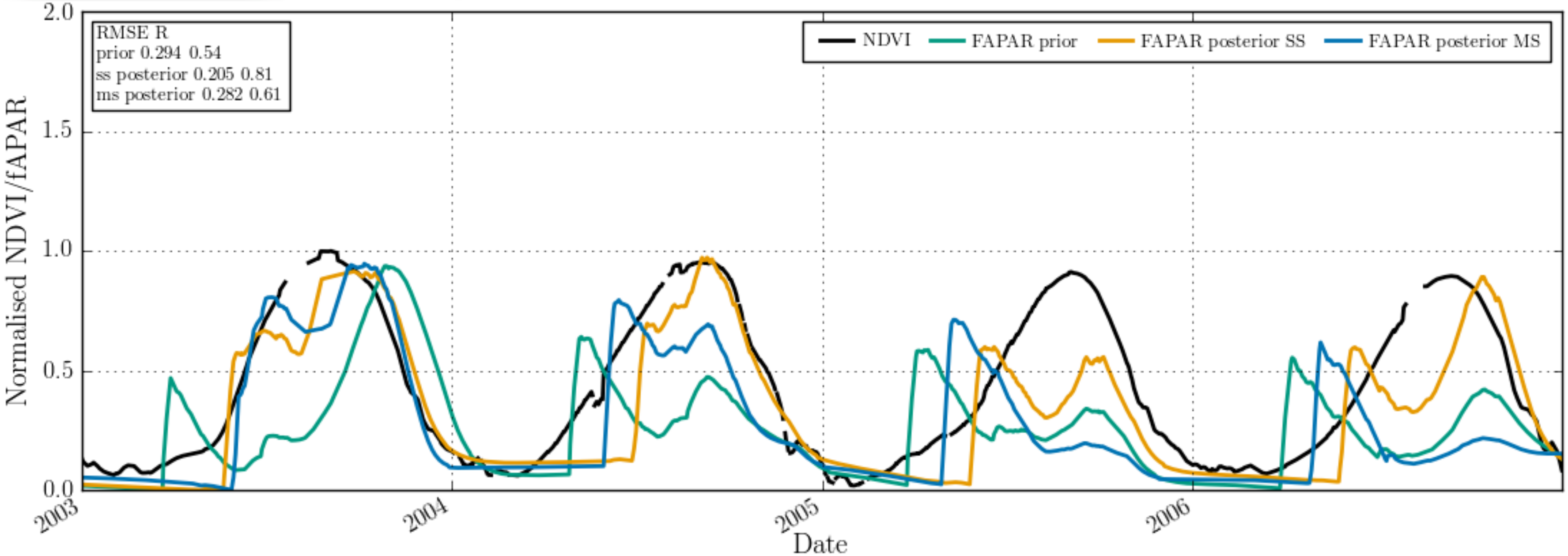


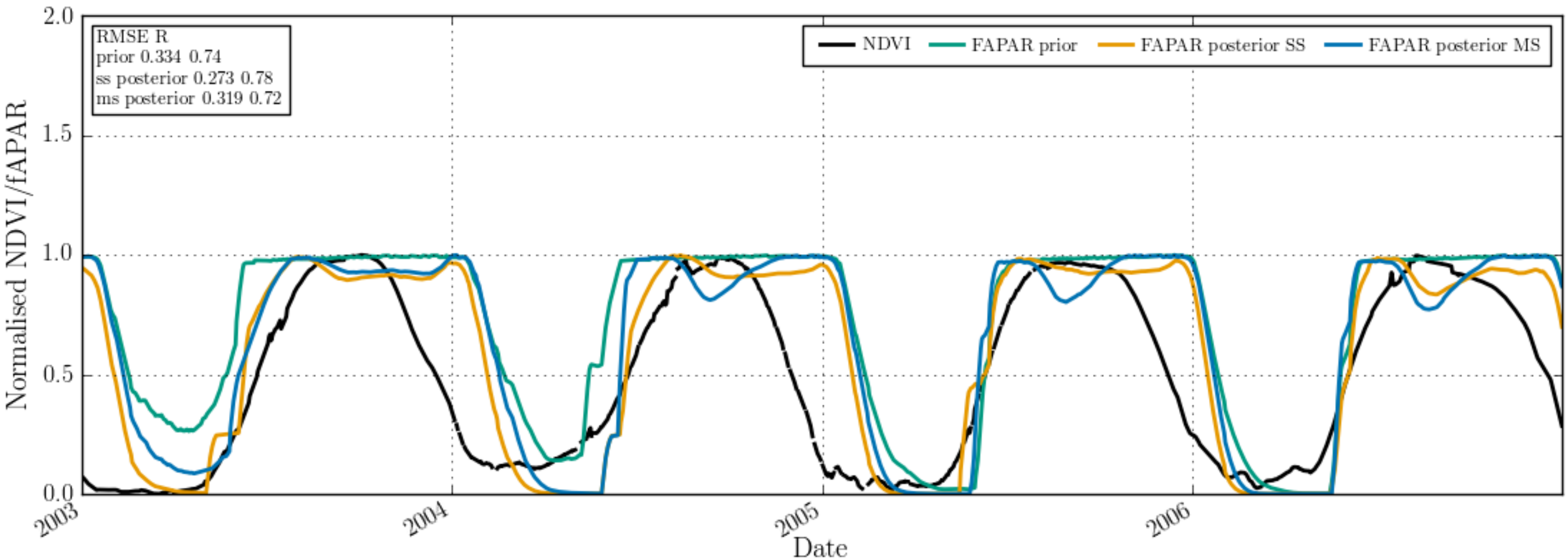
Tropical raingreen forest





Natural C4 grass





Spatial and temporal validation

- Multi-site posterior parameter used for validation

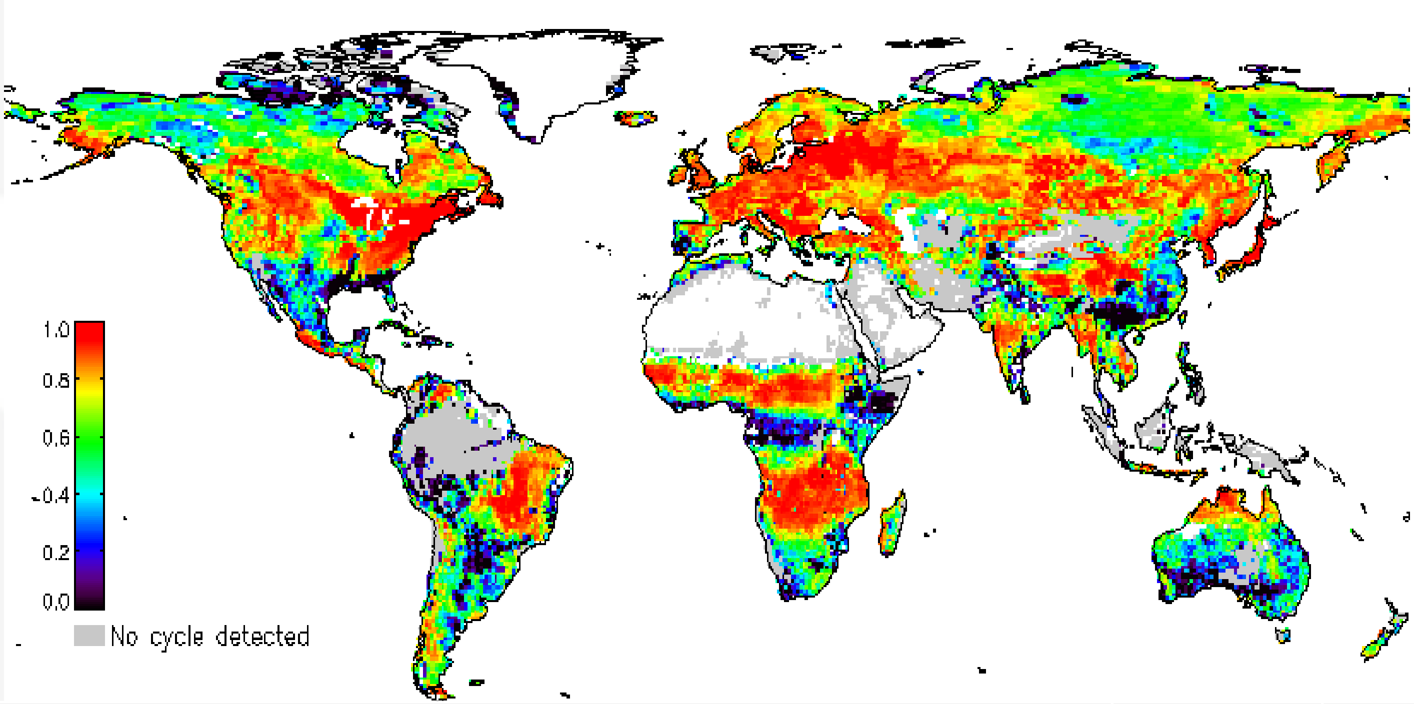
- Spatial validation
- Extra 15 grid points per PFT
- 2000 – 2008

PFT	Mean uncertainty reduction (%)	Prior Correlation	Posterior Correlation
TeBD	19	0.9	0.93
BoBs	13	0.59	0.65
BoNS	62	0.25	0.88
NatC3	24	0.63	0.74

- Temporal validation
- Original 15 optimisation grid pts
- Extra 2 years 2009 – 2010

PFT	Mean uncertainty reduction (%)	Prior Correlation	Posterior Correlation
TeBD	18	0.91	0.93
BoBs	28	0.55	0.72
BoNS	47	0.16	0.85
NatC3	24	0.6	0.75

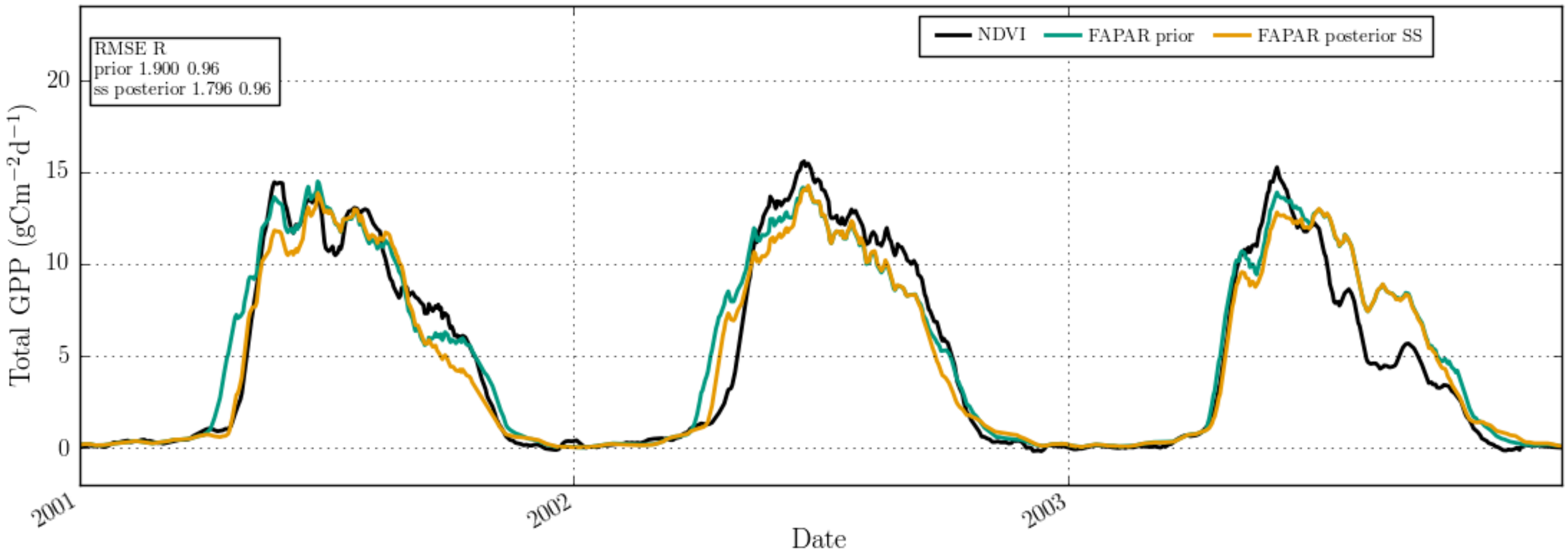
Global MODIS NDVI evaluation

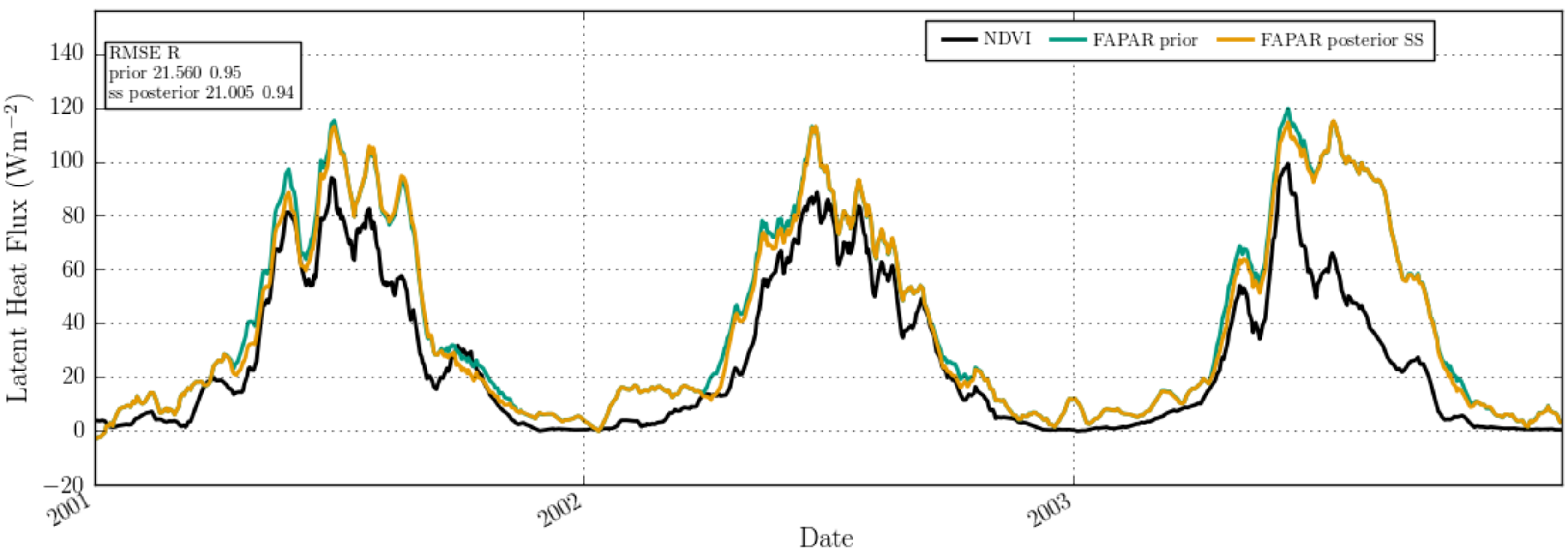


Median correlation value	prior	post1
PFT 6 temperate broad-leaved summergreen	**0.88**	**0.89**
PFT 8 boreal broad-leaved summergreen	**0.54**	**0.53**
PFT 9 boreal needeleaf summergreen	**0.36**	**0.91**
PFT 10 C3 grass	**0.53**	**0.59**

FluxNet evaluation - TeBS

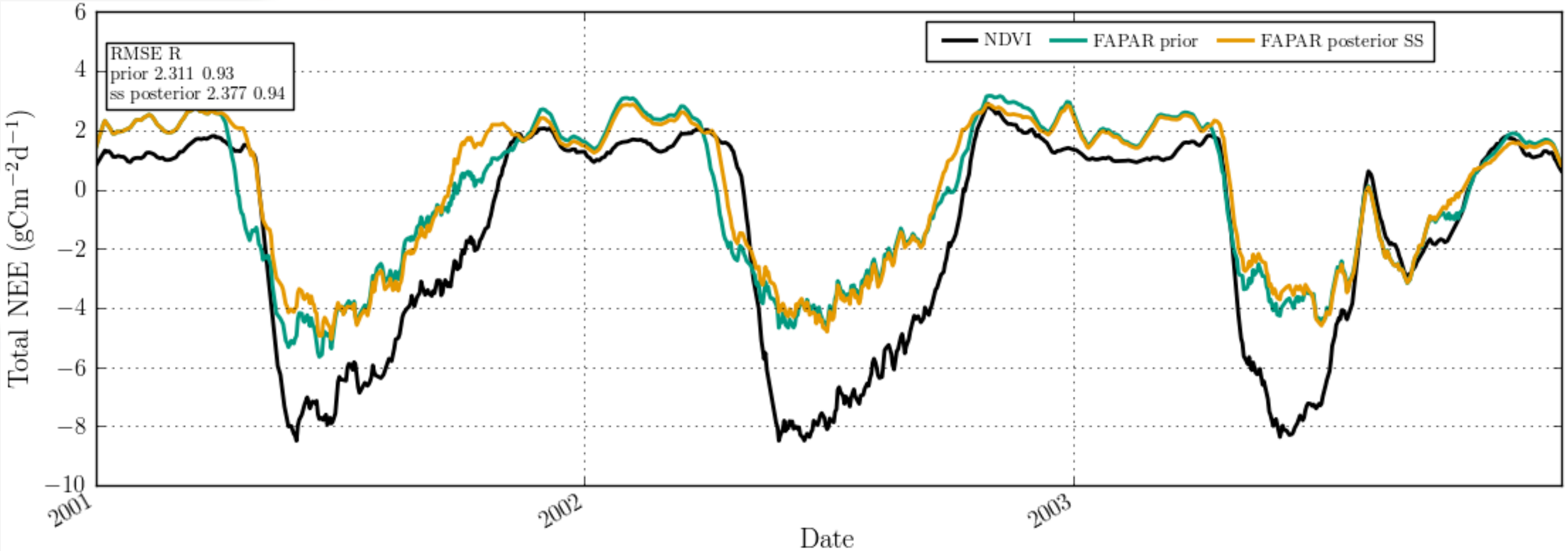
France (Hesse)

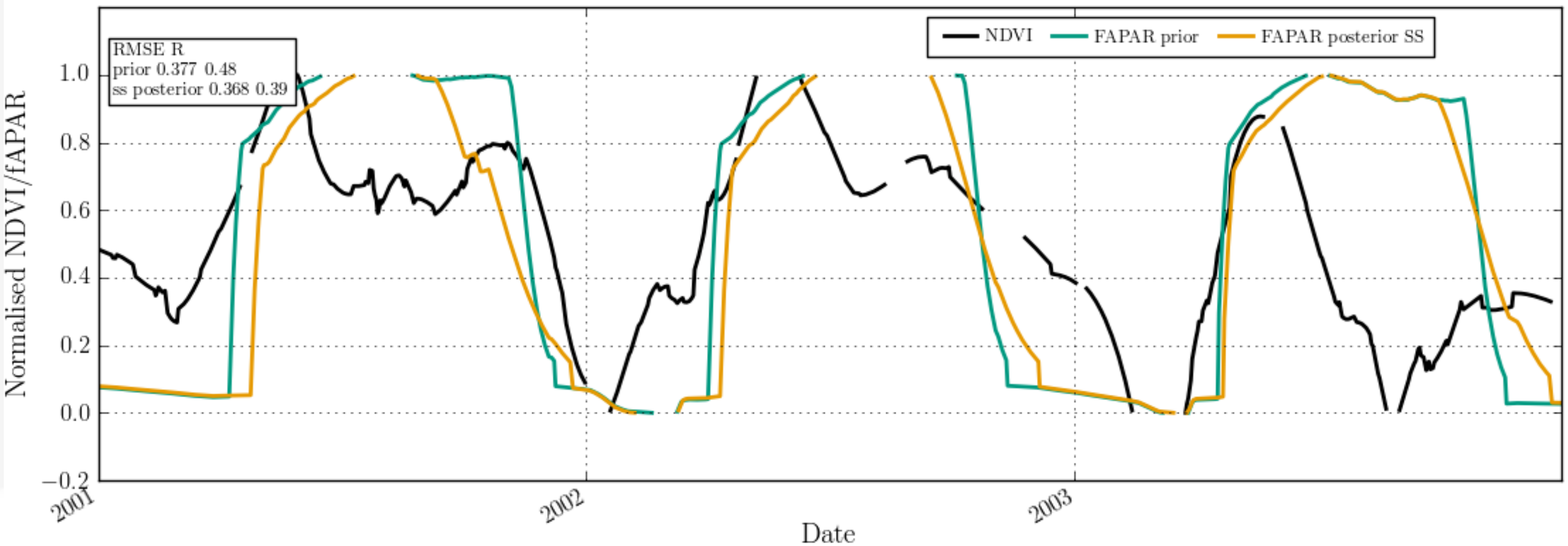




FluxNet evaluation - TeBD

France (Hesse)





FluxNet evaluation

TeBS:

- GPP → 10% mean reduction in RMSE
- NEE → -1%
- LE → 4%
- GSL mean bias (obs – model): -4 → -6

BoBS:

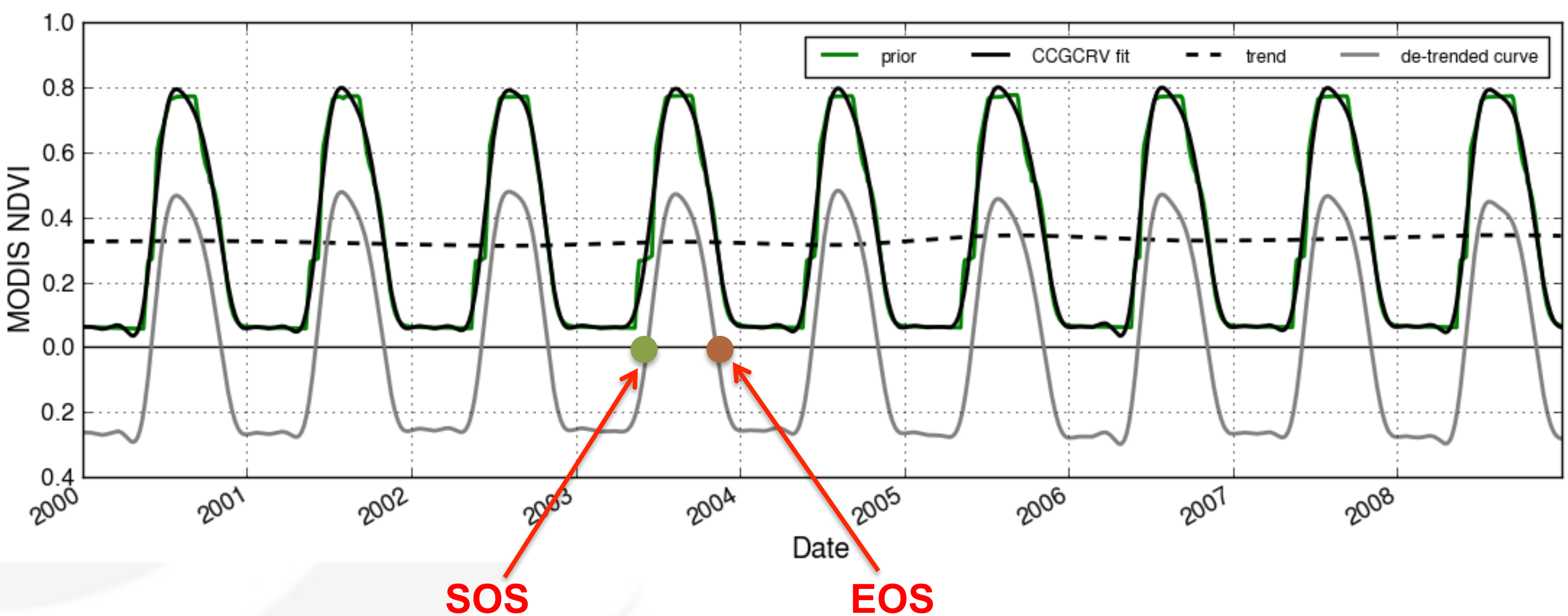
- GPP → 5% reduction in RMSE
- NEE → 5%
- LE → 4%
- GSL bias (obs – model): 9 → 46

NatC3:

- GPP → -5% mean reduction in RMSE
- NEE → -3%
- LE → -4%
- GSL mean bias (obs – model): -29 → -4

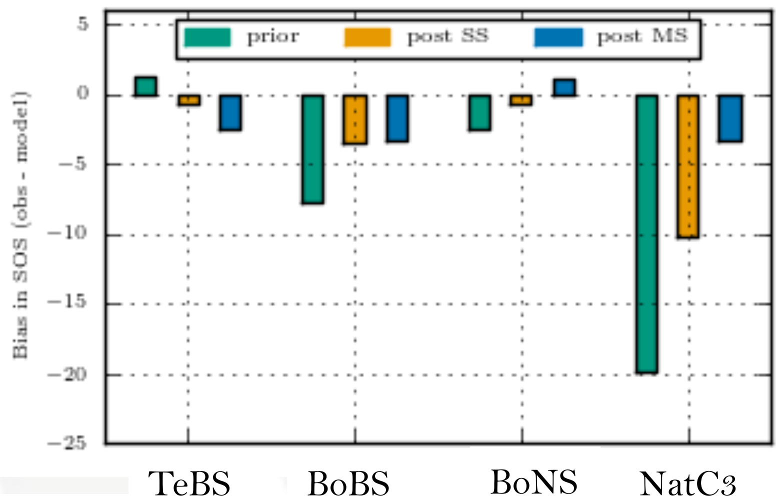
Impact on phenology metrics

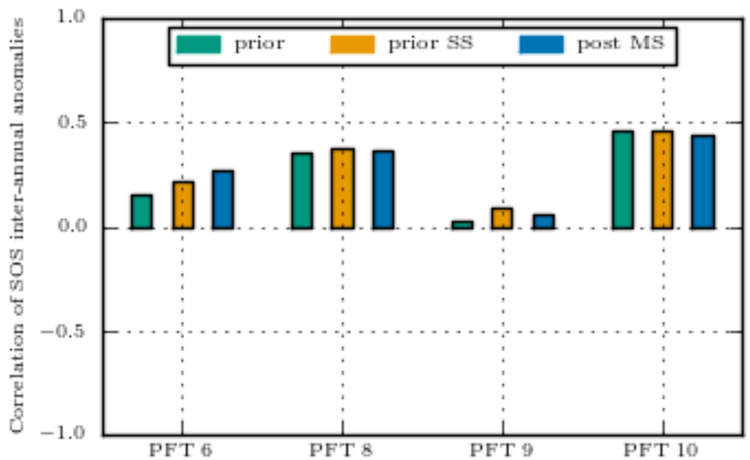
- CCGCRV curve fit (Thoning et al., 1989) → Fit and de-trend the signal
- Start of Season (SOS) and End of Season (EOS) when de-trended cycle crosses "zero line"
- Growing Season Length (GSL) = EOS – SOS



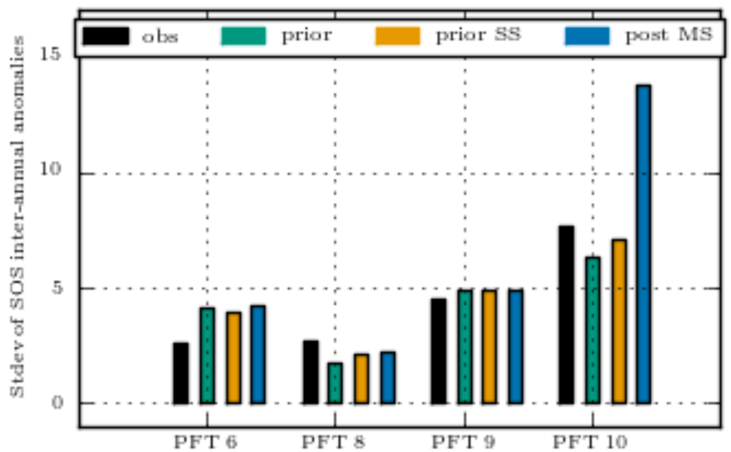
Impact on phenology metrics - SOS

Bias (obs – model)





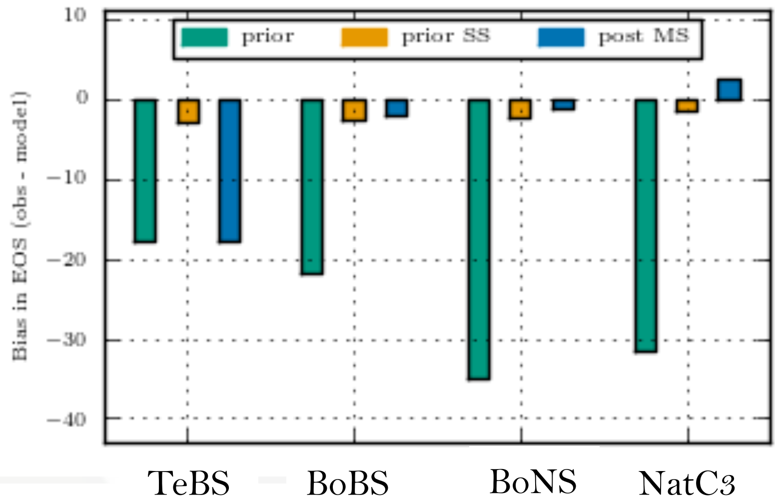
Correlations (btw model and obs) of inter-annual anomalies in SOS

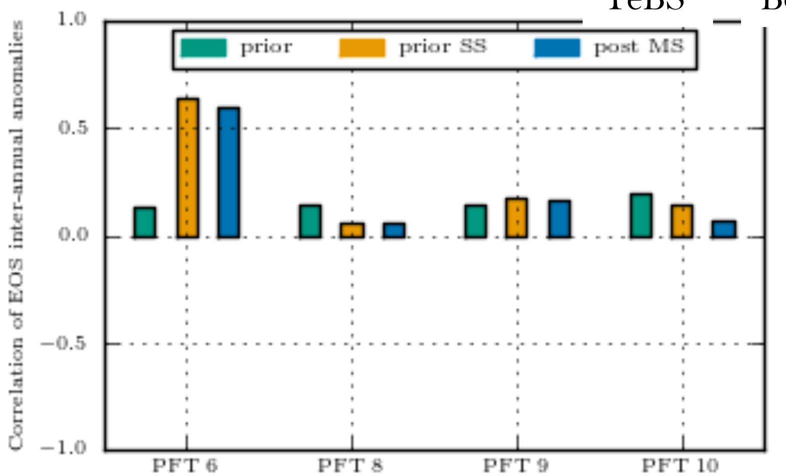


Stdev of inter-annual anomalies in SOS

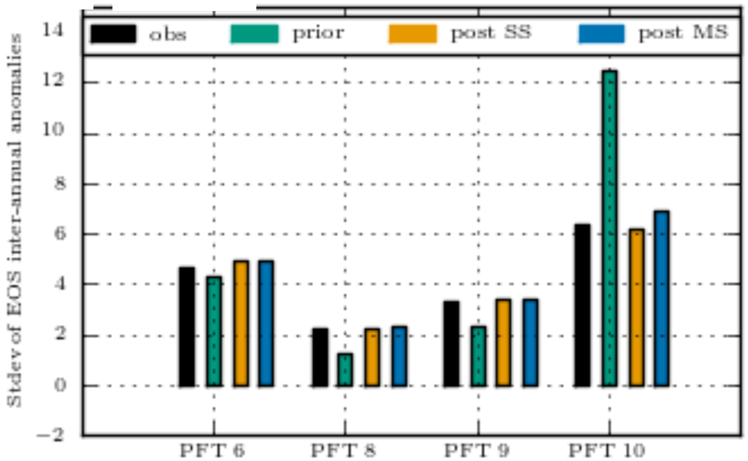
Impact on phenology metrics - EOS

Bias (obs – model)



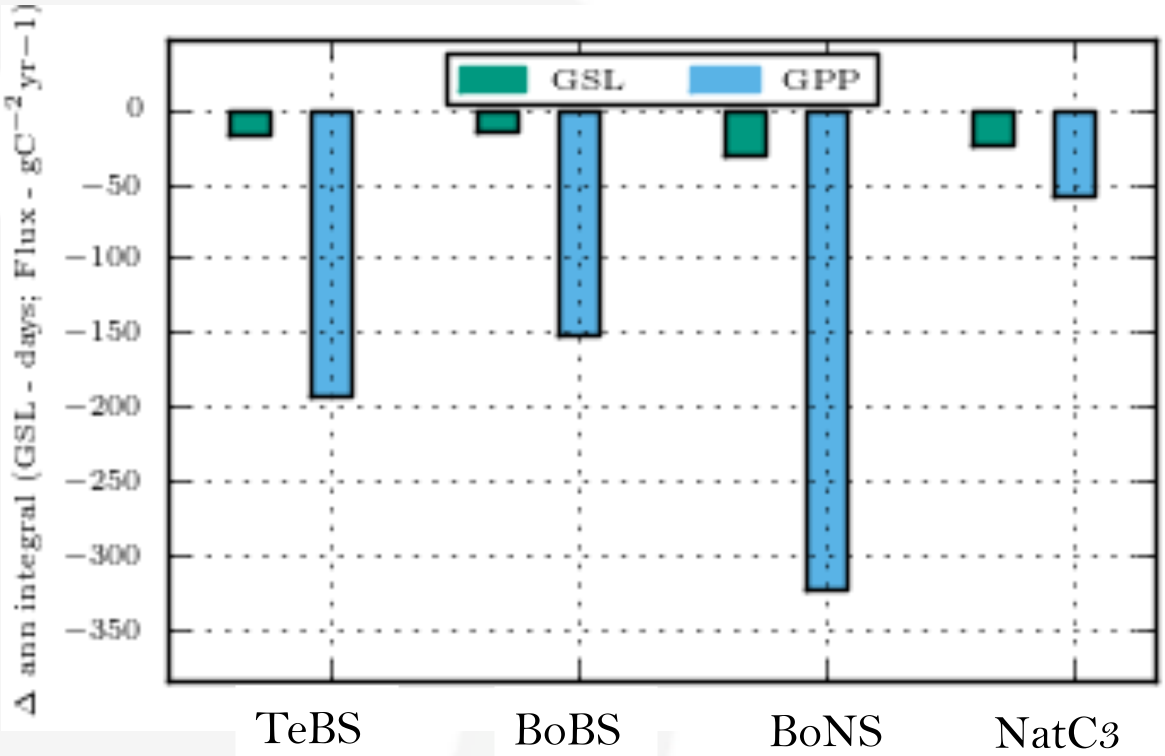


Correlations (btw model and obs) of inter-annual anomalies in EOS



Stdev of inter-annual anomalies in EOS

Impact of Δ GSL on net C fluxes



- PFT 6 : ~10gCm^{-2} / day
- PFT 8 : ~2.5gCm^{-2} / day
- PFT 9 : ~10gCm^{-2} / day
- PFT 10 : ~4gCm^{-2} / day

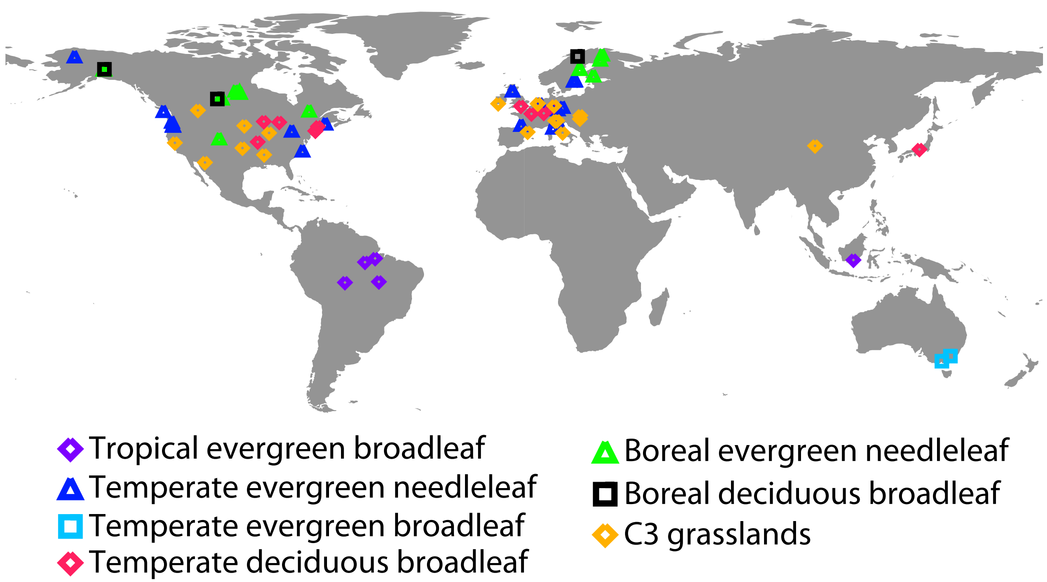
Summary of phenology optimisation

- Improved fit to satellite NDVI for temperate and boreal deciduous forest and grass (C3) *after* optimisation
- Reduction in GSL → earlier senescence → reduction in annGPP
- Improved fit to SOS. EOS harder to represent, *despite* main improvement in autumn
- Need for better understanding of PFTs where phenology driven by moisture conditions (tropical regions)
- Need to analyse impact on hydrology and energy budgets
- Move towards more PFTs or more generalised phenology model?

Further questions

- Questions of scale?
- Satellite versus in-situ data?
- Optimising mixed pixels?
- Normalising the data?
- Other data streams?

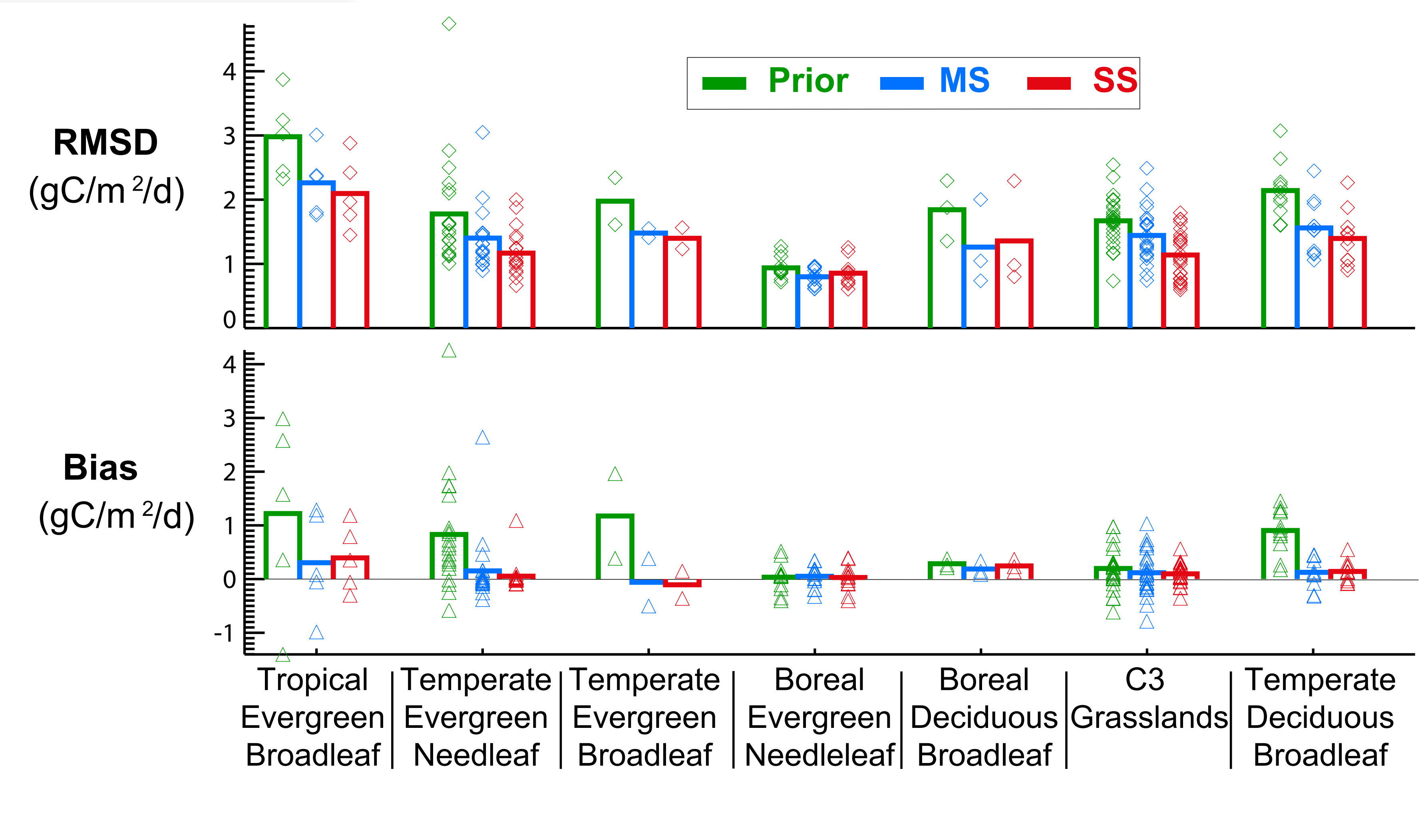
Fluxnet multi-site optimisations



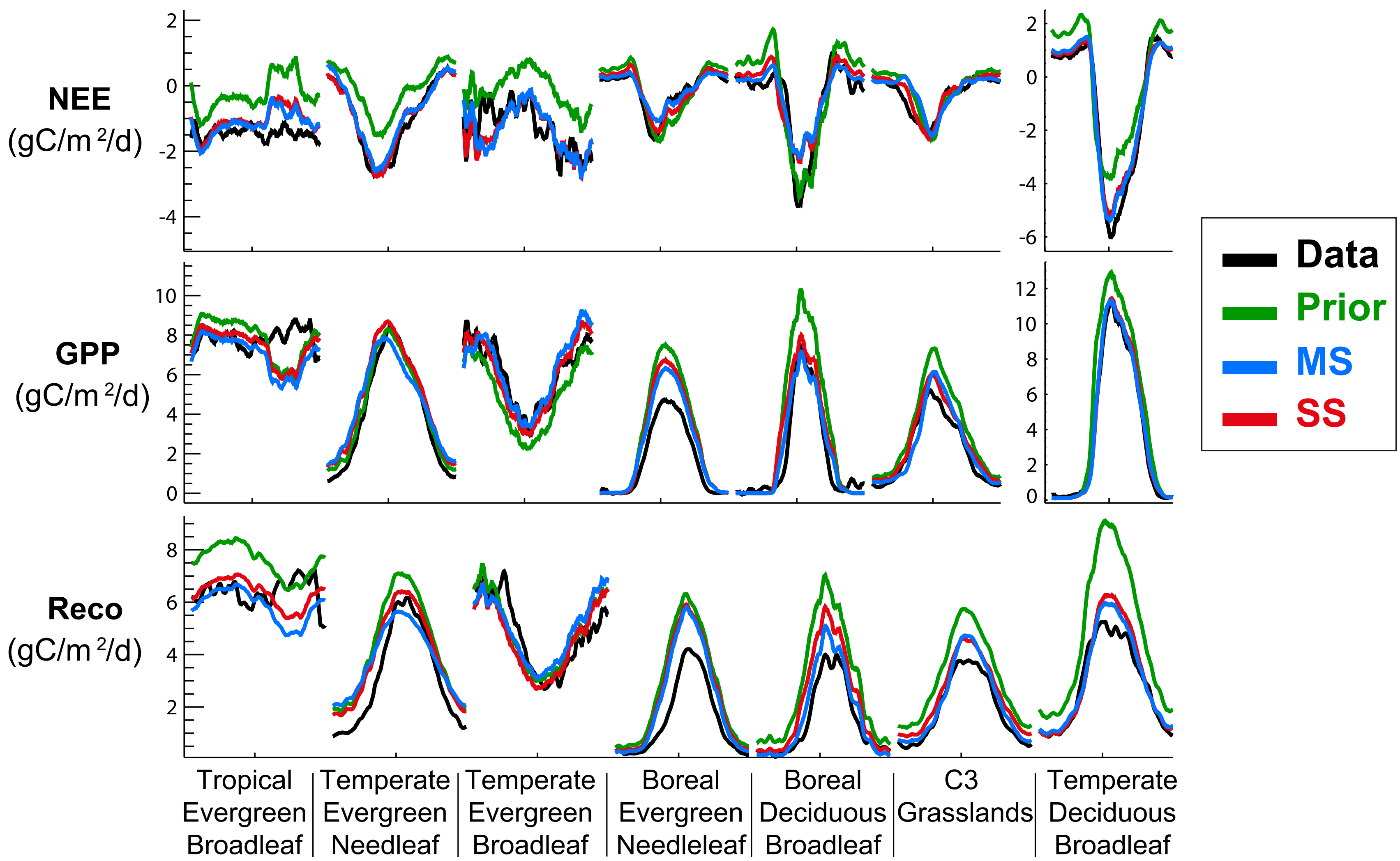
Parameter	Genericity
$V_{cmax,opt}$ $c_{T,min/opt/max}$ $L_{age,crit}$, $f_{stressh}$	PFT
$G_{s,slope}$	PFT
LAI_{MAX}, SLA	PFT
LAI_{init}	**Site**
$K_{lai,alloc}$	PFT
$K_{phenocrit}$, c_{senes}	PFT
MR_a, MR_b, GR_{frac}	PFT
Q_{10}, HR_b, HR_c Z_{decomp}	PFT
K_{soilC}	**Site**
$K_{albedo,veg}$	PFT

- Work done by Sylvain Kuppel during his PhD
- Figures taken from his soutenance presentation
- Refs: Kuppel et al. (2012) BG; Kuppel et al. (2014 - sub)

Fluxnet multi-site optimisations

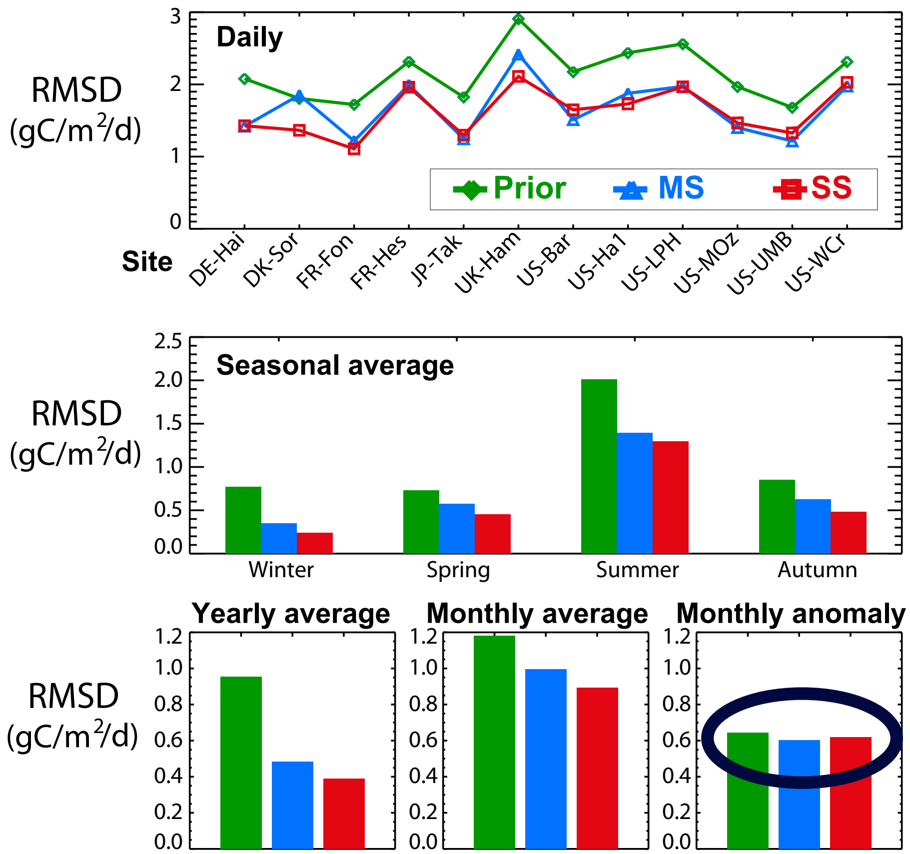


Fluxnet multi-site optimisations

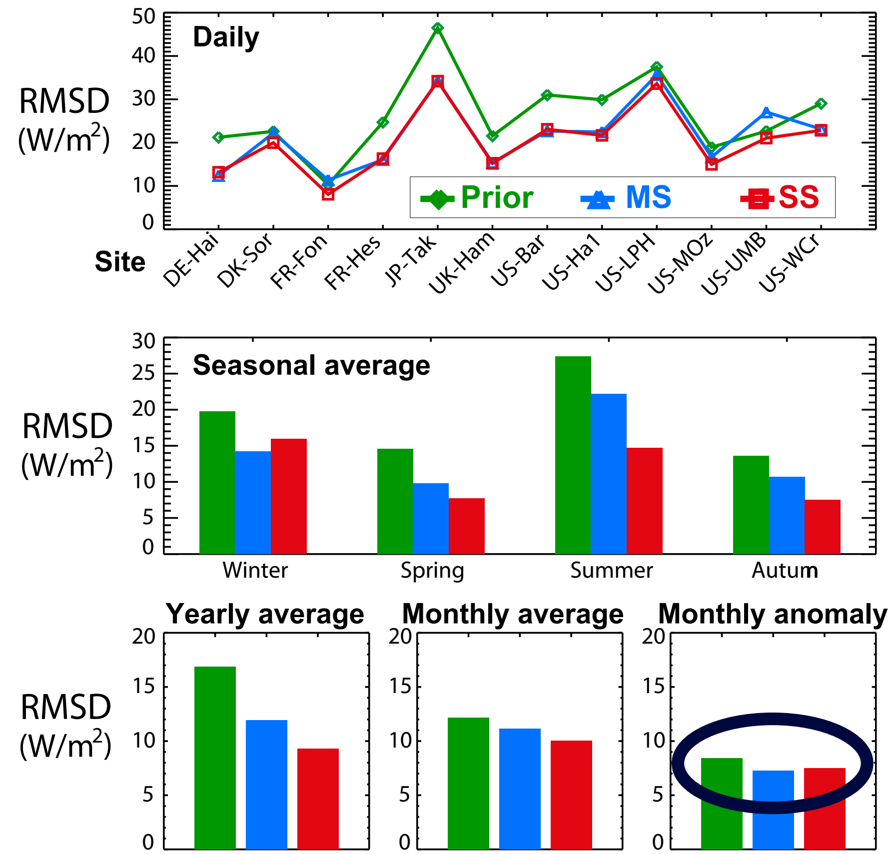


Improvement at different time scales

Carbon flux (NEE)

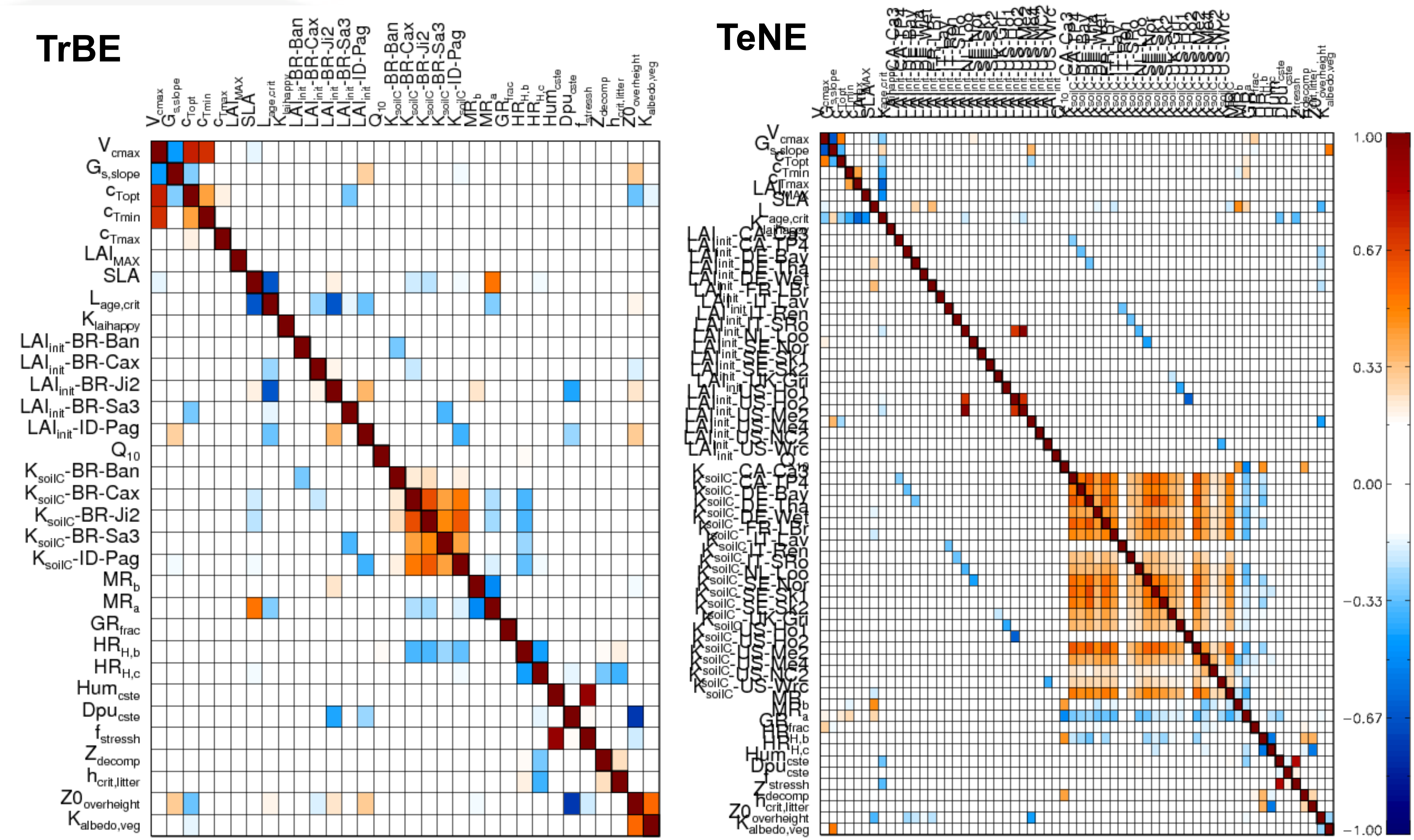


Latent heat flux (LE)

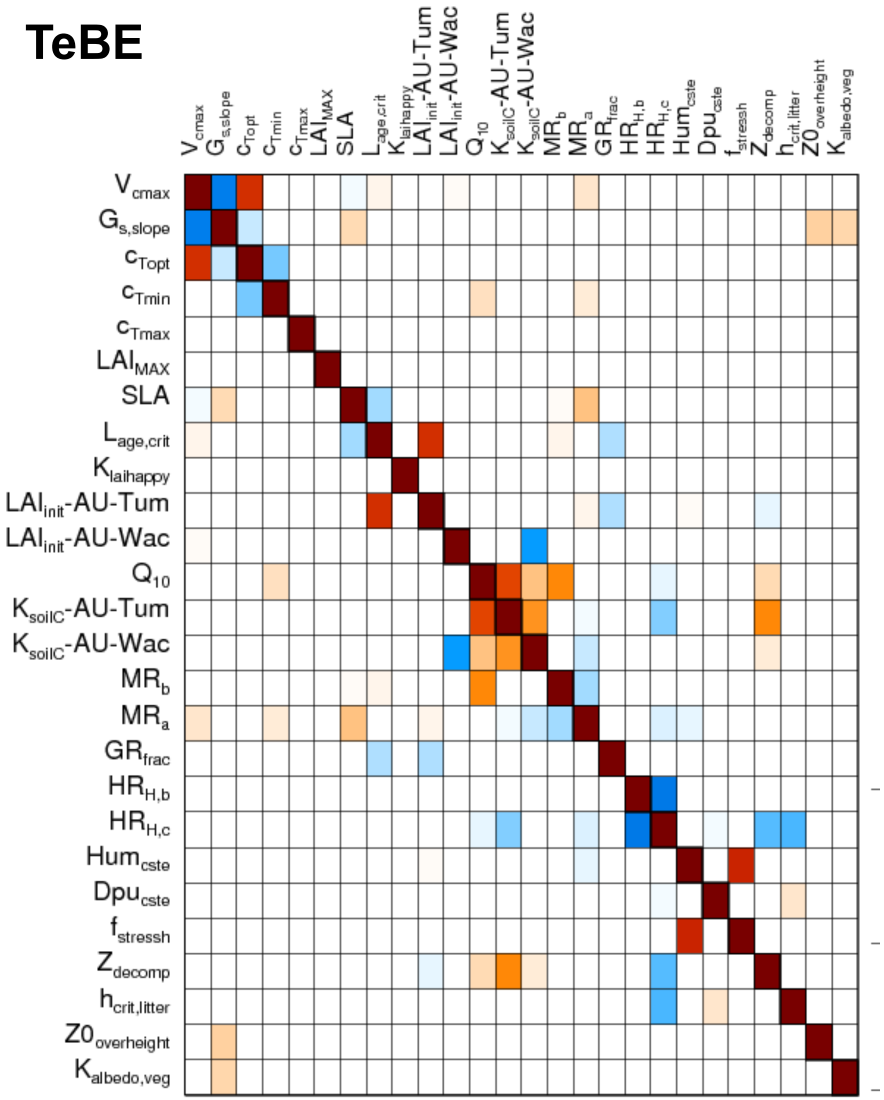


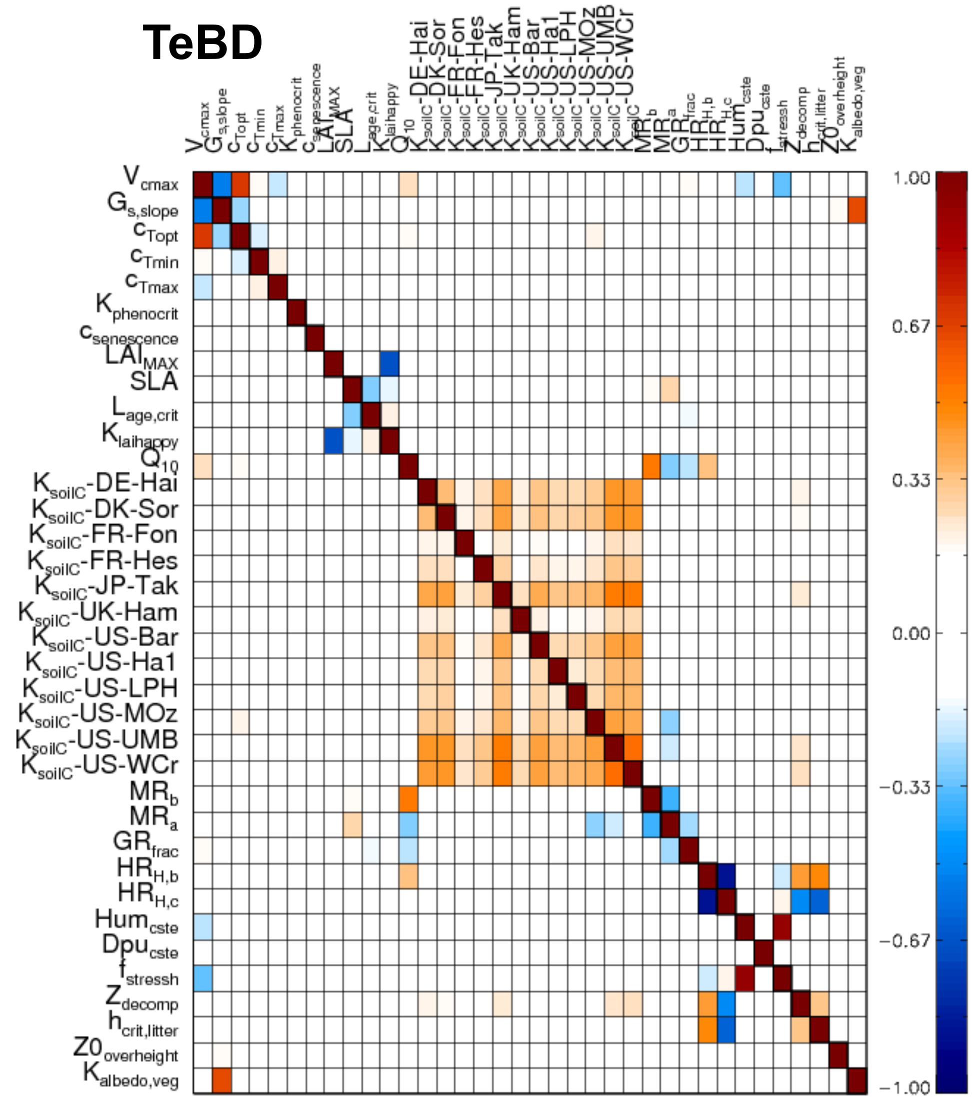
- Largest improvement of NEE at yearly time scale
- Similar performances between single-site (SS) and multi-site (MS)
- Small improvement of interannual flux variability

Parameter correlations

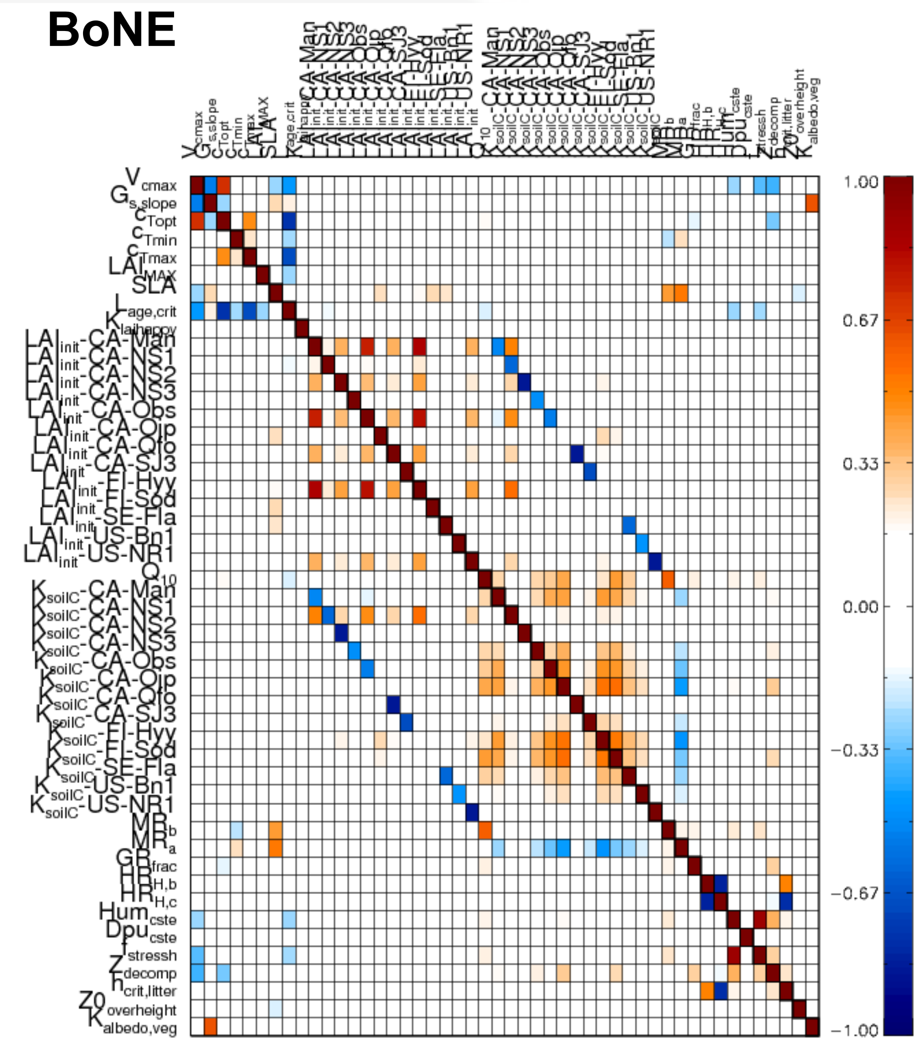


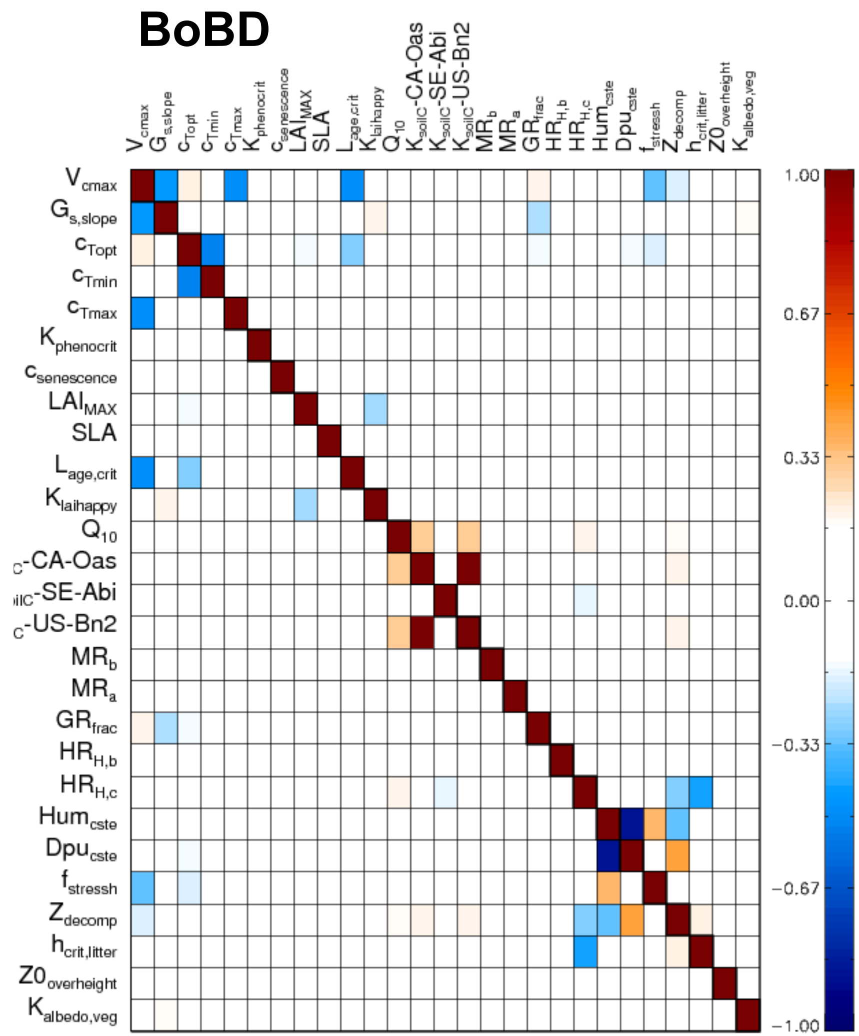
Parameter correlations



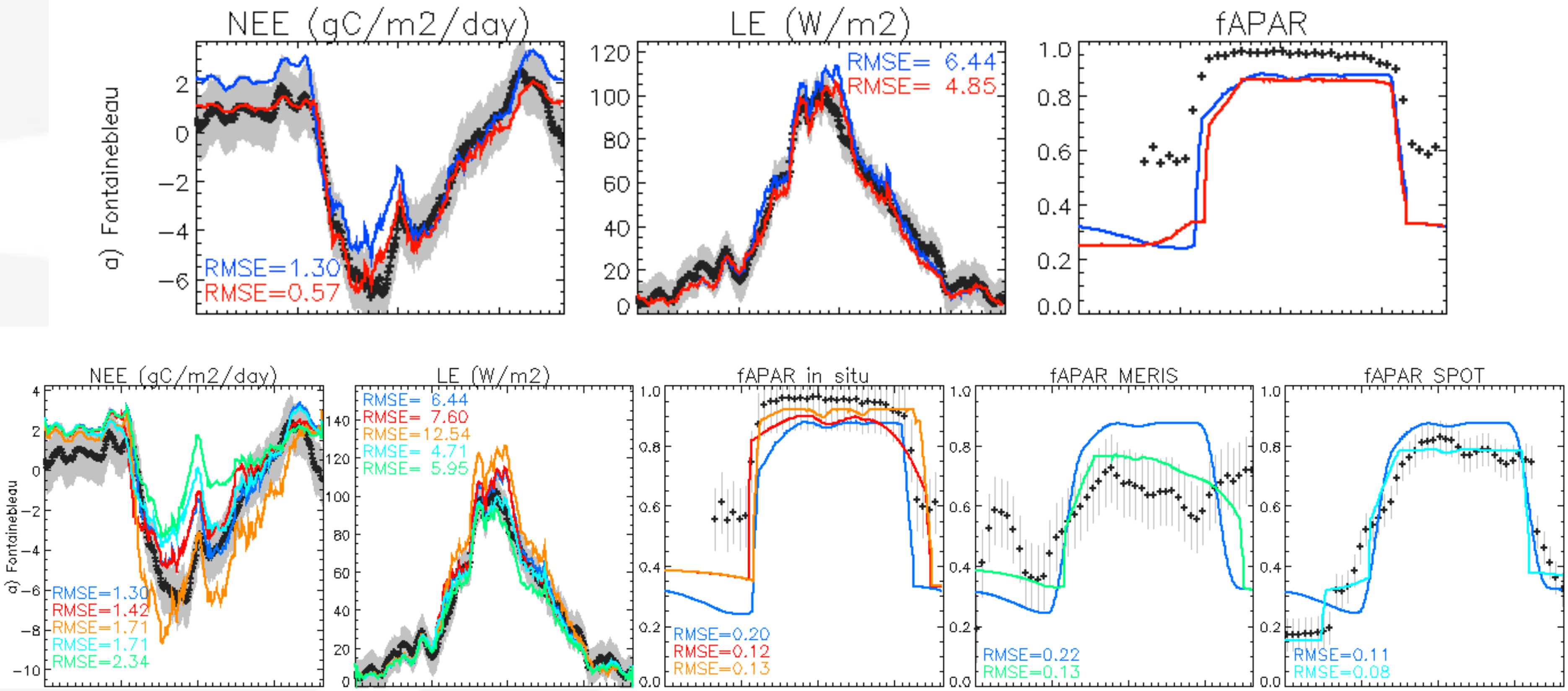


Parameter correlations





Importance of multiple data streams



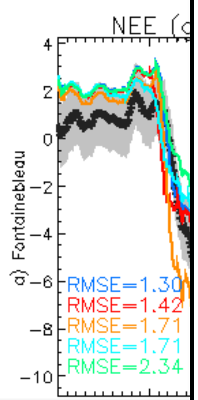
→ Also consider in-situ versus satellite fAPAR data

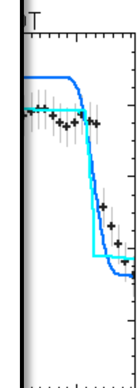
obs
prior
post fA in situ
post fA_ext in situ
post fAPAR SPOT
post fAPAR MERIS

Importance of multiple data streams

NEE (gC/m2/day) LE (W/m2) fAPAR

a) Fontainebleau





"This suggests the model does not find the CO_2 and fAPAR observations consistent. So rather than optimize parameters of the model, we have falsified some element of its structure.

Far from seeing this as a disappointment I would argue it is an exemplary application of data assimilation.

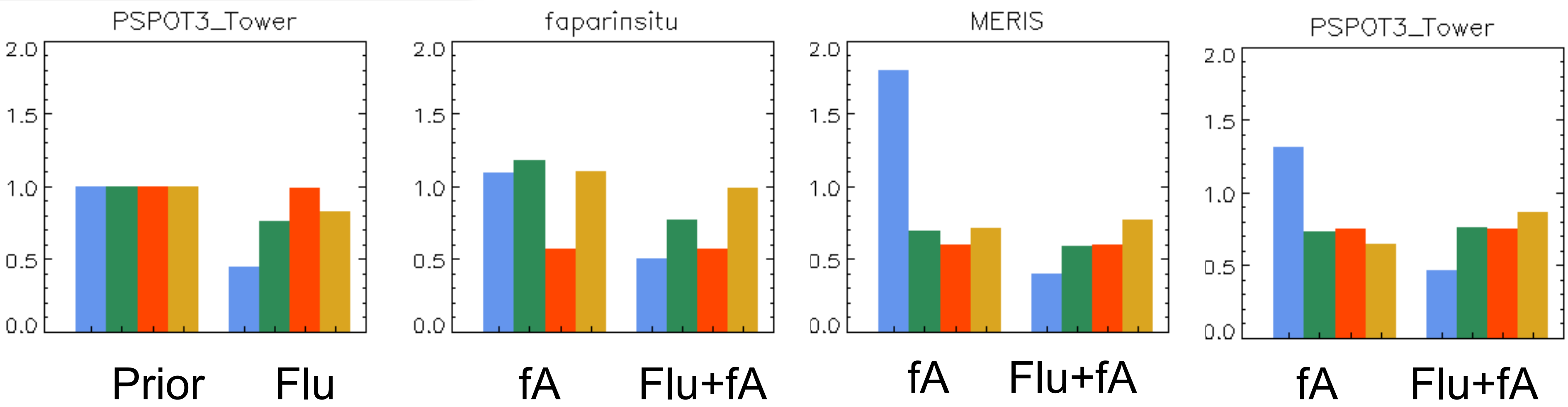
Note that if we had not carried out the parameter optimization we could never have distinguished between parametric and structural errors in the model."

Rayner P. (2010), The Current State of Carbon Cycle Data Assimilation, *Current Opinion in Environmental Sustainability*, **2**, 289-296

→ Also consider in-situ versus satellite fAPAR data

prior
post fA in situ
post fA_ext in situ
post fAPAR SPOT
post fAPAR MERIS

Importance of multiple data streams



Ratio between the posterior RMSE of fit and the prior RMSE, between the model simulations and the different observations:

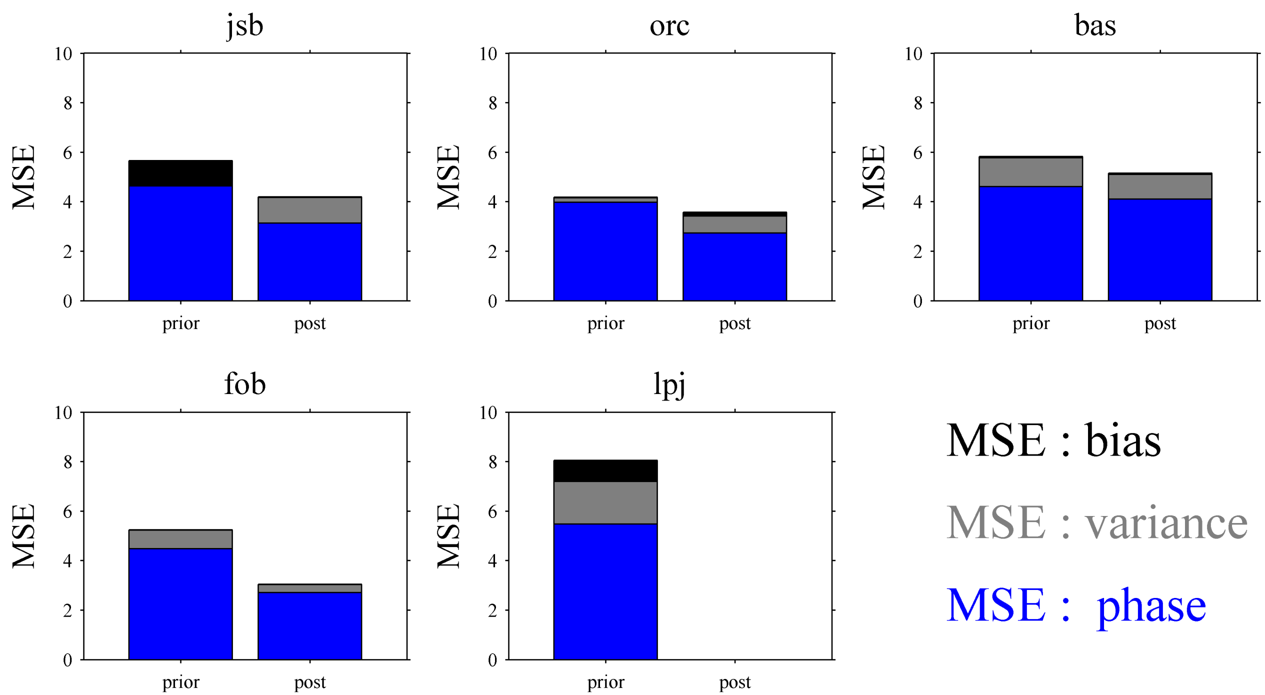
- **assimilations performed with only flux data (Flu),**
- **only fAPAR data (fA)**
- **combination of the two datastream (Flu+fA).**

Values < 1 (> 1) indicates model improvement (degradation).

NEE
LE
fAPAR
GPP

DA Intercomparison study – Fluxes

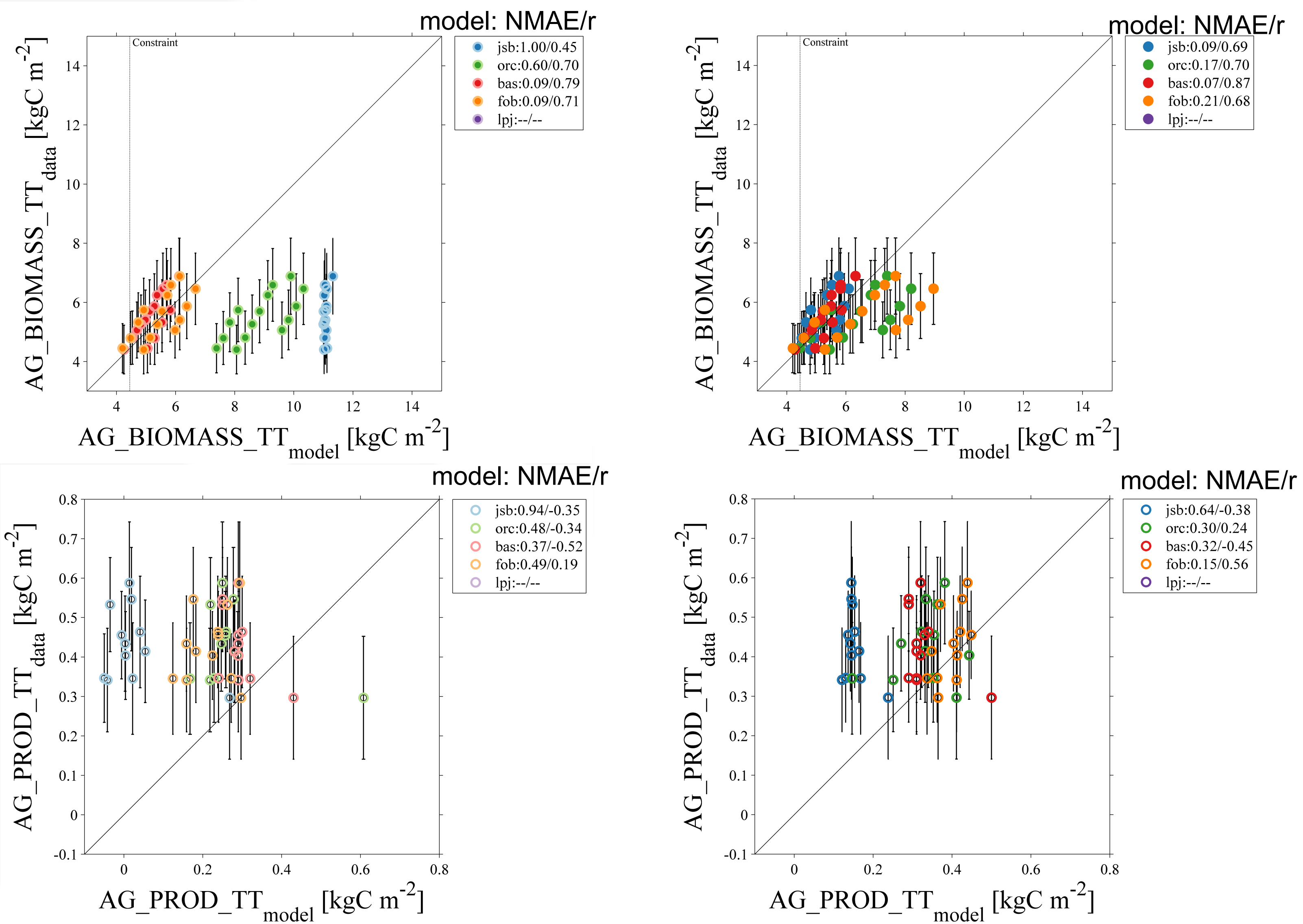
NEE at Hesse, France



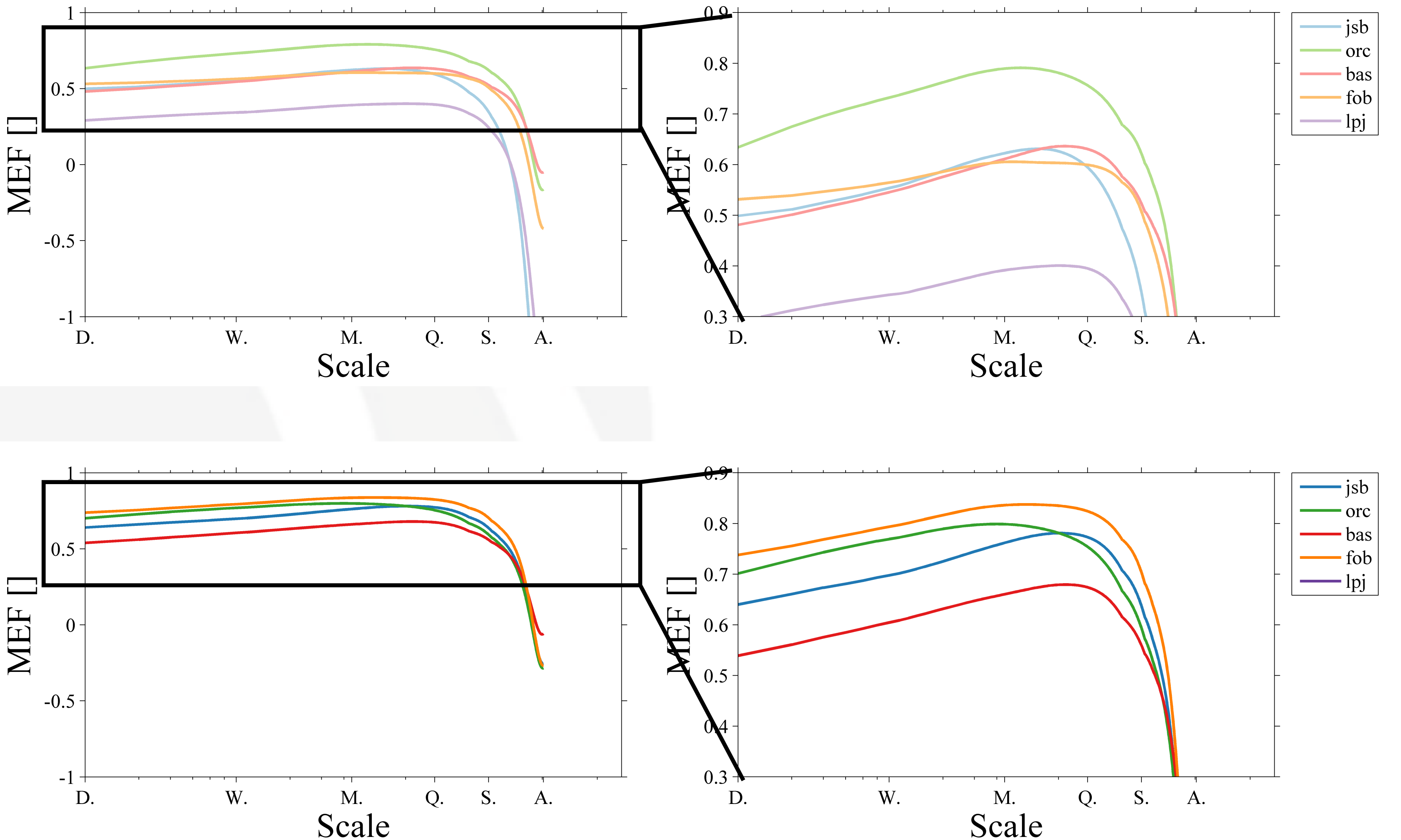
$$MSE = \overline{(M-D)^2} = 2\sigma_M \sigma_D (1-R) + (\sigma_M - \sigma_D)^2 + (\overline{M} - \overline{D})^2$$

$$= phase + variance + bias$$

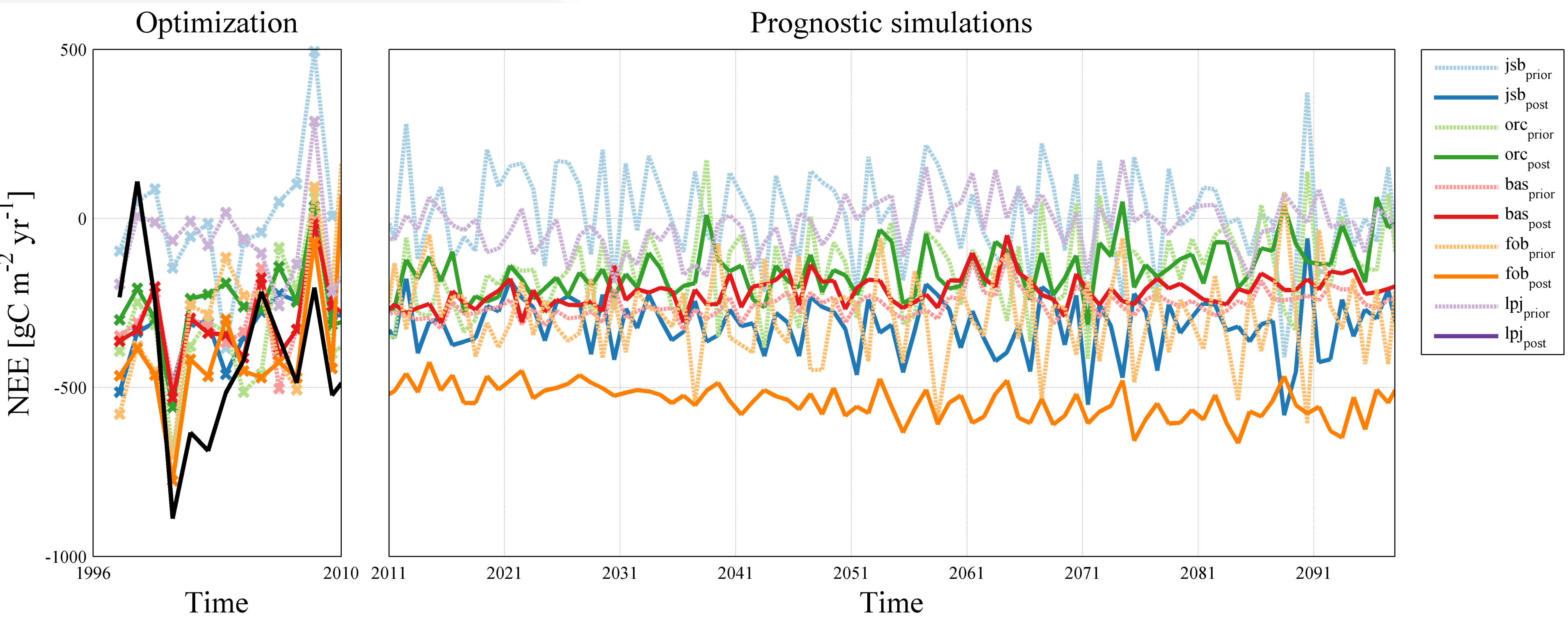
DA Intercomparison study – aboveground biomass



DA Intercomparison study



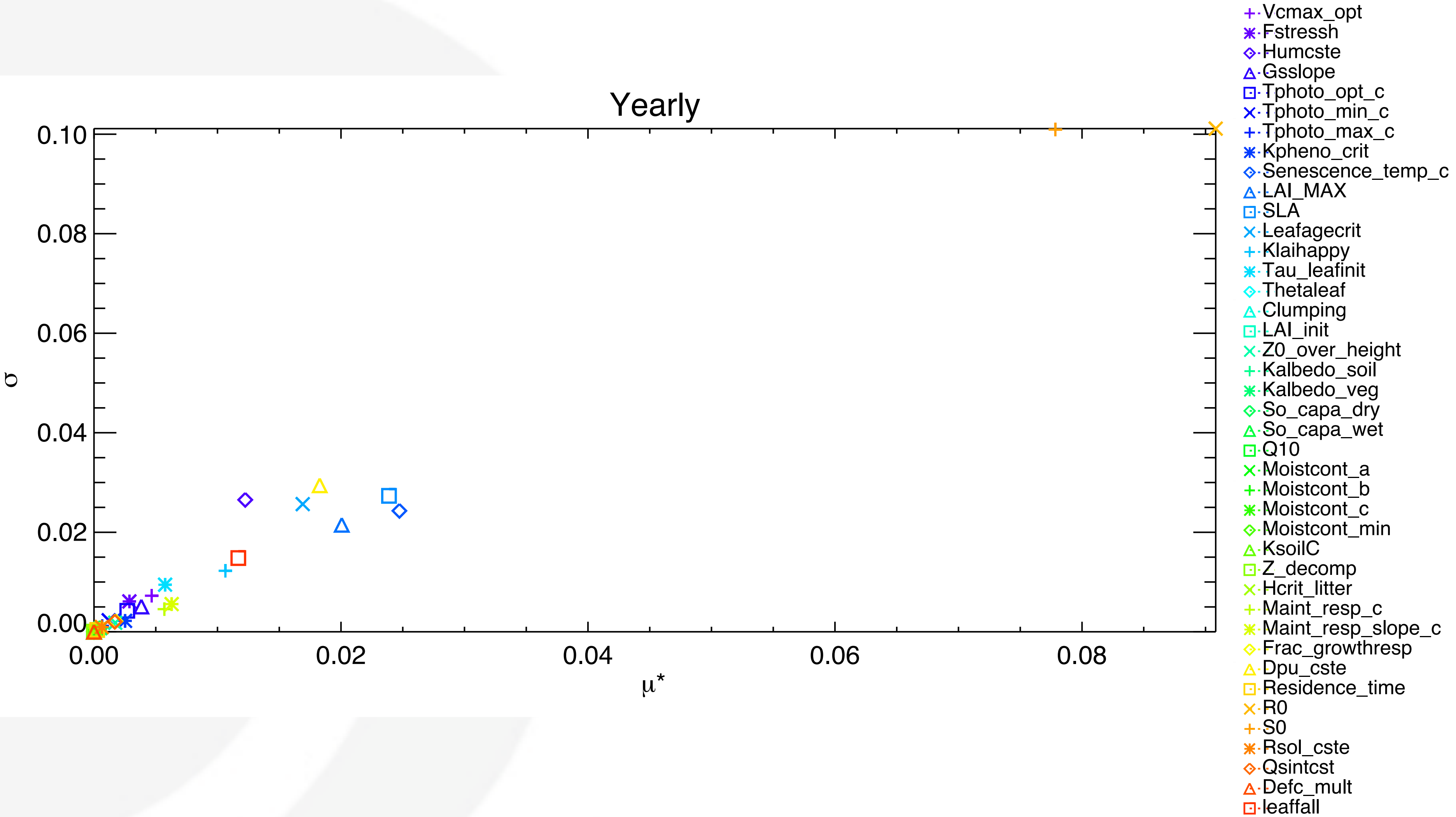
DA Intercomparison study – model spread



Summary

- DA SHOULD NOT JUST BE A BLACK BOX TOOL…
- Questions of scale?
- Optimising mixed pixels?
- Generality of posterior parameters / parameter correlations
- Model physics not accounted for?
- Do we have the right things for the wrong reasons?
- Importance of multiple data streams
- Interannual variability, partitioning of fluxes etc
- STILL WORK TO BE DONE…

Sensitivity of fAPAR – BoBS



Sensitivity of fAPAR – Natural C3