ExpeER
Ecosystem Research

ExpeER
Distributed Infrastructure for EXPErimentation in Ecosystem Research

Grant Agreement Number: 262060

SEVENTH FRAMEWORK PROGRAMME

CAPACITIES
INTEGRATING ACTIVITIES: NETWORKS OF RESEARCH INFRASTRUCTURES (RIS)
THEME: ENVIRONMENT AND EARTH SCIENCES

DELIVERABLE D10.2

Deliverable title: Report on upscaling frameworks and support for the infrastructure and model users

Abstract: This research describes different upscaling frameworks of local measurements made by ecosystem infrastructure (such as water and carbon fluxes), to larger scales like catchments or continents. Principles of the methods are first presented and practical implementation for two cases are then detailed: upscaling of water budget for a catchment area and upscaling of carbon fluxes and stocks at continental scales.

Due date of deliverable: Month 48

Actual submission date: Month 49

Start date of the project: December 1st, 2010

Duration: 54 months

Organisation name of lead contractor: CNRS

Contributors: Philippe Peylin, Natasha MacBean, Vladislav Bastrikov, Fabienne Maignan, Harrie-Jan Hendricks Franssen, Hanna Post

Revision N°: FINAL

Dissemination level:

PU Public (must be available on the website)	**[X]**
PP Restricted to other programme participants (including the Commission Services)	**[ ]**
RE Restricted to a group specified by the consortium (including the Commission Services) (precise to whom it should be addressed)	**[ ]**
CO Confidential, only for members of the consortium (including the Commission Services)	**[ ]**

Table of Content

1. EXECUTIVE SUMMARY 1

2. TEMPORAL AND SPATIAL UPSCALING: PRINCIPLES 2

2.1. FROM LOCAL OBSERVATIONS TO LARGE SCALE ESTIMATES 2

2.2. UPSCALING APPROACHES FOR BIOGEOCHEMICAL CYCLES 3

2.2.1. Parameter optimization 3

2.2.2. Model state variable optimization 5

3. UPSCALING OF WATER FLUXES AT THE BASIN SCALE 6

3.1. DATA REQUIREMENTS 6

3.2. UPSCALING FRAMEWORK BASED ON CLM 7

4. UPSCALING CARBON FLUXES AND STOCKS AT CONTINENTAL SCALE 16

4.1. HOW TO OPTIMIZE A PROCESS-BASED MODEL: EXAMPLE WITH ORCHIDEE 16

4.1.1. Choosing observations 16

4.1.2. Selecting model parameters 19

4.1.3. Finding the optimal parameters and their uncertainties 22

4.1.4. The need to assimilate multiple sites 24

4.1.5. How to combine multiple data streams? 27

4.2. USING AN OPTIMIZED MODEL TO UPSCALE CARBON FLUXES AND STOCKS 30

Upscale quantities and their uncertainties 30

Evaluation of the optimized model 30

5. CONCLUSION / NEXT STEPS 31

6. REFERENCES 33

1. Executive summary

This reports describes methods for upscaling of local measurements made by ecosystem infrastructure, such as net ecosystem exchange (NEE) or latent heat fluxes (LE) to large catchments or continents. Rather than an exhaustive literature survey of possible upscaling approaches, we describe a general framework based on the assimilation of local observations in a process based ecosystem model and detail two different implementations based on model state variables or parameters optimization.

Parameter or model state optimizations are usually formalized with the Bayes theorem that describes how to optimally combine observations and models and to derive optimal parameters/variables with the help of measurements. With the hypothesis of Gaussian errors (observations, parameters or model states) the optimization is equivalent to the minimization of a cost function describing the model data misfit and the departure to prior knowledge on model parameters/states. Different methods that are classically used to solve this problem are described, from sequential approaches (i.e., Ensemble Kalman Filters) to Variational approaches.

A framework for water flux upscalling for the Rur catchment in Germany is proposed with the CLM model. The method requires high quality meteorological forcing, information on land use and plant functional type and soil texture maps to calculate hydraulic properties. A first example is provided with the optimization of both model parameters and model states at the Rollesbroich grassland site in the Rur Catchment using soil moisture time series at 41 locations and 3 depths. Different optimization methods are tested; they highlight a significant improvement of model soil moisture for a “verification period” using the parameters optimized for the assimilation window. The feasibility of a joint model state and model parameter optimization is illustrated but the difficulties linked to spatially varying hydraulic properties are discussed. An upscalling at the Rur catchment with the use of satellite surface temperature at 1 km resolution (from MODIS) to optimize few model state variables and parameters is presented. The improvement of the model latent heat and sensible heat fluxes at few sites remains small which highlights the difficulties of the approach at the basin scale.

A second framework for carbon flux upscaling at the continental scale with the optimization of the ORCHIDEE model parameters is then proposed. Different steps or issues that are usually encountered when upscaling carbon fluxes at large scale with process-based model are detailed. They concern i) the choice of observations with their information content, ii) the selection of the most sensitive parameters of the model (illustrated with a MORIS sensitivity analysis), iii) the difficulties to find the optimal parameter set linked to the level of model non linearity (strengths and weaknesses of variational versus monte carlo methods are discussed), iv) the need to assimilate measurements from multiple sites in order to obtain generic parameters (discussion of multiple-site versus single-site optimizations with eddy covariance NEE and LE fluxes) and v) the difficulties to combine data streams of different nature such as carbon fluxes and stocks (for forest). The potential of the method based on the assimilation of various data streams is then summarized.

2. Temporal and spatial upscaling: principles

2.1. From local observations to large scale estimates

We consider in this project an up-scaling framework as a modelling approach to estimate ecosystem properties such as water and carbon fluxes or stocks at large spatial and/or temporal scales using primarily local-scale observations. We will focus on the prediction of water and carbon fluxes for a basin or a continent using an ensemble of in-situ flux/stocks measurements. Different frameworks exist to upscale in space and time depending on the type of model, the use or not of satellite observations and the diversity of local-scale observations that are combined. We can distinguish two different conceptual frameworks although many variant exist:

- ”Data oriented” approaches: In these approaches the observations acquired at local scale are directly valorized using statistical relationships between the target quantities (fluxes and/or stocks) and explanatory variables that i) can be used to estimate these quantities and ii) can be measured at large scale (regional to continental for space; present to future for time). These approaches often use satellite observations, such as vegetation activity indexes (NDVI, fAPAR) or surface water and temperature observations, for the spatial upscaling. The approach of Jung et al. 2011 is one of the most cited one to upscale globally carbon fluxes from in-situ flux measurements.

- ”Process-based model” oriented approaches: This method relies on the use of a process-base ecosystem model that describes, through an ensemble of equations, the different processes controlling the flow of carbon and water in the soil-plant-atmosphere continuum. The upscaling framework is now defined as a way to combine the information from the model and from local observations, given that the model drivers (meteorology, vegetation and soil descriptions) are available at large scale. Different ways of using the local observations to improve the model estimates can be implemented, from model state to parameters optimization. Figure 1 summarizes a common approach where i) model parameters are optimized at local scale using an ensemble of observations and ii) the optimized model is then run at large scale.

In this report we focus on the second type of approach. We first describe the principles and two different approaches; we then discuss more in details practical implementation for water balance upscaling at the basin scale and carbon flux upscaling at continental scale. Note that the upscaling of biodiversity is discussed in the deliverable D10.1.

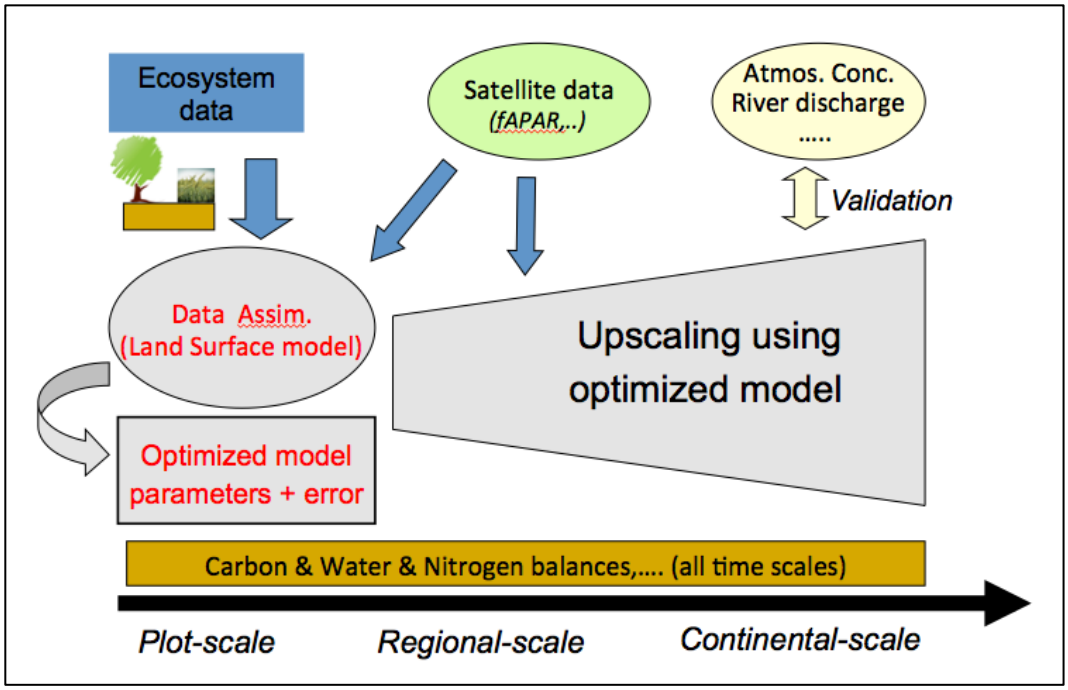


Figure 1: diagram detailing a typical up scaling framework (in space and time) with the optimization of a land surface model using in situ or satellite observations of the carbon, water and energy cycles. The optimized model is then run at larger scale with the optimized parameters in order to estimate "optimal" water and carbon fluxes.

2.2. Upscaling approaches for biogeochemical cycles

Two different approaches are detailed below, the optimization of model parameters and the optimization of model state variables.

2.2.1. Parameter optimization

Land surface model simulations are affected by several uncertainty sources which influence the quality of the characterization or prediction. The most important uncertainty sources include the model forcings (e.g., precipitation, short wave radiation), model parameters (e.g., rooting depth, the maximum rate of carboxylation, soil hydraulic conductivity and many others), initial conditions (e.g., carbon and nitrogen pools) and not in the last place model structural error. Model structural error refers to an incomplete understanding of the governing processes. Examples are the empirical description of soil respiration or the use of the Richards equation to model flow in the unsaturated zone on large spatial scales. The relative importance of the different error sources depends on the application. For example, for weather prediction the uncertainty of the initial conditions is believed to be the most important one and therefore ensemble predictions focus on the characterization of this source of uncertainty, while other uncertainty sources are neglected/maintained deterministic. Another example is groundwater flow where uncertainty of the hydraulic conductivity is thought to be the most important one and other uncertainty sources are often neglected.

Here we are concerned with predictions of water, energy and carbon exchange between the land and the atmosphere and a complication is that all mentioned uncertainty sources are

relatively important to take into account. Parameter uncertainty is important, which is recognized already longer in carbon modelling studies (e.g., Raupach et al., 2005; Braswell et al., 2005), and only recently has been taken into account in land surface modelling studies with a focus on hydrology (e.g., Pauwels et al., 2009; Han et al., 2014). Uncertainty of initial conditions is especially important for carbon studies as the magnitude of the carbon pools might deviate strongly from the steady-state model assumption, for example related to management practices (e.g., Carvalhais et al., 2008). However, this source of uncertainty is often neglected in many carbon modelling studies. Forcing error is commonly recognized as an important source of uncertainty in calculations with land surface models and treated as a random variable in most studies. Finally, model structural error is important for land surface models as many processes, especially those ones concerning the vegetation are described in a rather empirical manner in the models. In this work, model structural error was neglected, and it was considered that model forcings, parameters and initial conditions were the main sources of uncertainty. A focus was put on parameter estimation (and less on estimation of initial conditions), and the uncertainty of model forcings was treated as a random component.

Measurements offer the possibility to improve predictions with the imperfect land surface models and reduce the influence of the different sources of uncertainty. It is however important to take measurement uncertainty into account. Data might show complex error structures, measured net ecosystem exchange (NEE) by the eddy covariance method (EC) for example has a measurement error which depends on the flux magnitude and is distributed double exponentially (Richardson et al., 2006; Hollinger and Richardson, 2005).

In order to improve prediction and characterization with land surface models, it is important to optimally combine measured data and simulations with land surface models. Bayes law offers a framework to correct model predictions with help of measurements:

$$p(\theta|\tilde{Y}) = \frac{p(\tilde{Y}|\theta)p(\theta)}{p(\tilde{Y})}$$

where ϑ are modelled states and/or parameters, Y are data, $p(\theta|\tilde{Y})$ is the posterior probability density function of states and/or parameters, $p(\tilde{Y}|\theta)$ the likelihood function, $p(\theta)$ the prior distribution of states and/or parameters and $p(\tilde{Y})$ the normalizing constant of the given data. An exact solution of Bayes law for arbitrary probability density functions is not available, and with Monte Carlo type techniques approximations can be calculated for such cases. Assuming a Gaussian distribution of states and parameters, the following objective function can be derived:

$$J(x) = \frac{1}{2}[(H(x) - y).\mathbf{R}^{-1}.(H(x) - y)^T + (x - x_b).\mathbf{P}_b^{-1}(x - x_b)^T]$$

where the first term in brackets measures the misfit between measurement data and simulated values and the second term the deviation with respect to the prior estimates of the initial states and/or parameters. Here y is a vector containing the measurements, x a vector containing the states and/or parameters, x_b a similar vector as x, but now containing the initial estimates for the states and/or parameters and H is an operator that allows a

comparison of the modelled and measured values at the measurement locations. **R** represents the error variance/covariance matrix associated to the observations (gathering both the measurement and model errors) and $\mathbf{P_b}$ the parameter error variance/covariance matrix. This objective function is minimized and the typical formulation used in variational data assimilation. The formulation allows to evaluate a longer time series of measurement data/modelled states at once. In this work, variational data assimilation was applied to estimate unknown ecosystem parameters, and in some simulation experiments initial conditions were also uncertain, and treated as additional parameters to be estimated. As a summary, variational data assimilation was used as a parameter estimation framework. In section 4 more details can be found on this approach.

2.2.2. Model state variable optimization

Sequential data assimilation (SDA) techniques can also be derived from Bayes law. SDA incorporates measurement data for single time steps and modifies model states and/or parameters on the basis of measurement data at this single time step. A Markovian assumption is invoked. If also a Gaussian assumption is made for the error distribution of states, parameters and observations, the Kalman Filter formulation is obtained. The Monte Carlo variant of the Kalman Filter, the Ensemble Kalman Filter (EnKF) (Evensen, 1994; Burgers et al., 1998) is more robust for non-linear model dynamics. Traditionally, sequential data assimilation is used to update system states in real time, without parameter calibration. Land surface modelling with a focus on soil moisture/hydrology was traditionally oriented on updating of states alone. Typically, EnKF is used as updating equation. The main steps in EnKF are:

i) A large number of ensemble members is generated and for each of these ensemble members the forward equation (land surface model) is solved until a certain time step where observations are available that have to be assimilated in the model. The ensemble members are for example obtained by using different model forcings and different model parameters in each of the runs. Model forcings and model parametes are sampled from a prior probability density function.

ii) The predictions with the land surface models in step i) will be updated with help of measurements. Therefore measurements are compared with simulated values at the measurement locations. The measurement equation is given by:

$$y_i^k = y_i + w_i^k$$

where *y* is again the vector with observations and *w* is the vector with observation errors, which is typically generated from a normal distribution $N(0, \sigma)$ and where σ is the expected measurement standard deviation. The subscript *i* indicates the ensemble member and superscript *k* timestep.

iii) In the updating step the model forecasts from i) are corrected by the measurement data from ii). The Kalman gain is calculated which weights model simulation results and measurement data at each spatial location. The Kalman gain is given by:

$$\mathbf{K} = \mathbf{C}\mathbf{H}^T(\mathbf{H}\mathbf{C}\mathbf{H}^T + \mathbf{R})^{-1}$$

using similar symbols as before, and the matrix **C** contains the model covariances. The updating equation is then given by:

$$\mathbf{x}_i^{a.k} = \mathbf{x}_i^{f,k} + \mathbf{K}(\mathbf{y}_i^k - \mathbf{H}\mathbf{x}_i^{f,k})$$

where *f* refers to the forecast (before updating with measurements) and *a* to analysis (after the update with measurement data). This equation is evaluated to update each of the ensemble members.

Updating only state information allows to adjust model states towards measured values of the states and in case the system has a considerable memory, this improves the forecast over a significant time horizon. As soil moisture (especially the deeper soil moisture) has a memory which might extend over some months, updating the states alone should result in an improved characterization of soil moisture content over a longer time period. Updating states related to groundwater might even have a longer effect as groundwater reacts slowly on weather changes. This background was the motivation for the popularity of SDA, and EnKF in particular, in land surface hydrology, but as indicated before, recently there is an increased focus on estimating parameters together with the states.

3. Upscaling of water fluxes at the basin scale

3.1. data requirements

Upscaling of water and carbon fluxes at the basin scale requires a large amount of data as input. Whereas we focus on upscaling of carbon fluxes in section 4, here we focus on the upscaling of water fluxes using the Community Land Model (CLM, http://www.cgd.ucar.edu/tss/clm/). The quality of upscaling of water fluxes is strongly dependent on the quality of spatially distributed input data for the land surface model. The main sources of spatially distributed input data for the land surface model CLM are:

i) Meteorological forcings in a high temporal resolution. Often a temporal resolution of 30min is used. Input data are air temperature, humidity, wind speed, incoming shortwave and longwave radiation and air pressure.

ii) Information on land use and plant functional type. For the main application in the EXPEER project, the Rur catchment, the four main plant function types are cropland, grasland, evergreen needleleaf forest and deciduous broadleaf forest. CLM uses default parameters for plant functional types.

iii) Information on soils. Input on soil texture is used in CLM to calculate, with help of a pedotransfer function, soil hydraulic (and thermal) properties like saturated hydraulic conductivity, the B exponent of the Clapp-Hornberger equation and porosity.

For data assimilation purposes also information on measurements including measurement uncertainty is needed. For example, time series of soil moisture data and their uncertainty have to be provided. The frequency of assimilation has to be fixed beforehand.

For upscaling of hydrological fluxes we used a sequential data assimilation framework as outlined before. The Markovian assumption, which is made in sequential data assimilation, is less problematic for the characterization of hydrology than for the characterization of carbon fluxes. In the latter case, many parameters act over long timescales and assimilation in batch has the advantage that parameters are adjusted to fit the long time series.

3.2. Upscaling framework based on CLM

In contrast to upscaling of carbon fluxes, data assimilation for land surface models focusing on the hydrological cycle is more focused on state estimation. However, recently some papers also showed that parameter estimation might be important, especially for the improved characterization of evaporative fluxes (e.g., Han et al., 2014). We tested for the small Rollesbroich site, and later for the larger Rur catchment, whether parameter estimation with sequential data assimilation can improve the characterization of hydrological states. Estimation of hydrological relevant parameters with land surface models has hardly been evaluated with real data, therefore this test was performed. In this case, the state vector of EnKF is augmented and includes both states and parameters. Accordingly, the model covariance matrix includes correlations between model states and parameters.

Evaluation of parameter estimation at plot scale

The Rollesbroich grassland site (50°37'27"N, 6°18'17"E) is a subcatchment of the TERENO Rur catchment in Germany [Bogena et al., 2010], located in the Eifel. The soil texture is silty loam. It covers an area of 27 ha with an altitude ranging between 474 and 518 m.a.s.l. The mean air temperature is 7.7 °C, the mean precipitation is 1033mm, and the mean slope is 1.63°. At the site an eddy covariance tower (50°37'19"N, 6°18'15"E, height 514.7m.a.s.l) and a soil moisture and soil temperature sensor network (with measurements at 5, 20 and 50cm depth) are installed, amongst others. Soil moisture time series at 41 locations are being recorded.

The Rollesbroich site is modelled as a single point and the data of the soil sensor network are averaged over the complete area. An areal average of soil moisture content at 5cm, 20cm and 50cm depth was calculated. The forcing data in this study (hourly air temperature, air

pressure, relative humidity, wind speed, incoming short/longwave radiation), were measured at the eddy covariance tower. Precipitation was measured by a tipping bucket located close to the eddy covariance station. Soil texture was determined for the area based on 273 soil samples, taken from three different depths, ranging between 5 and 11 cm, 11 and 35 cm, and 35 to 65 cm. The sample locations coincided with the location of the SoilNet sensors. Finally, for each of the three depth ranges average values were calculated. For CLM, input soil data are sand fraction, clay fraction, organic matter density and soil color.

CLM was spun-up from 1 January 2011 to 29 February 2012. The assimilation period was from March 1, 2012 to July 31, 2012. Daily soil moisture observations were assimilated for this period to update model states and possibly also parameters. The verification period was from August 1, 2012 to December 31, 2012. In this period, models were not informed by observations, but used the updated parameter values as input.

Soil moisture content measured at 5cm, 20cm and 50cm depth was assimilated. In CLM, the 10 predefined soil layers were used for hydrological calculations. Given this division in layers in CLM, soil moisture content measurements at 5cm, 20cm and 50cm corresponded to the third, fifth and the sixth model layer in CLM. The parameters updated for the third layer were also used for the first and second layer, the updated parameters for the fifth layer were copied to the fourth layer and the updated parameters of the sixth layer were copied to layer 7 until 10.

Different simulation experiments were performed: (i) Open loop run. In this case no data were assimilated and a free run was performed for the different ensemble members for the period March 1, 2012 to December 31, 2012. (ii) Updating of states only. In this case, soil moisture was updated by the soil moisture observations for the observation period only. (iii) Updating of both states and parameters. In the assimilation period, soil moisture and some parameters were updated by soil moisture observations. The updated parameter values from the last time step of the assimilation period were used in the verification period. These experiments were not only performed for EnKF using the augmentation approach, but also for EnKF and parameter updating according a dual approach (e.g., Moradkhani et al., 2005a), the sequential importance resampling particle filter (SIR-PF) (Moradkhani et al., 2005b) and the Markov Chain Monte Carlo particle filter (McMC-PF) (Vrugt et al., 2013). The results for the algorithms differed only marginally, with slightly better results for the EnKF-algorithms. Results will not be shown here. Finally, the land surface model VIC-3L was also tested for the same site and the same data assimilation algorithms and showed in general a slightly worse performance compared to CLM. These results are also not presented here.

100 ensemble members were used in the open loop and data assimilation experiments. Precipitation and soil parameters were perturbed to present the uncertainties in the model simulation. Precipitation (mm) data were multiplied by (1+0.1* random sample from standard normal distribution). The small perturbation of precipitation data is motivated by the available on-site measurements. Further information on the perturbation of the soil parameters is given in Table 1. The soil moisture observation error was assumed normally distributed with mean equal to 0 and standard deviation equal to $0.02 m^3/m^3$. The model error was set to zero assuming that uncertainty was captured by uncertain parameters and model forcings. Parameter inflation according to Whitaker and Hamill (2012) was applied.

The prior ensemble standard deviations for the parameters were used to maintain ensemble spread. The standard deviations of the parameters directly follow from the distributions used to perturb the initial parameter values.

Results for the assimilation period show that Nash-Sutcliffe values (NSE) and root mean square errors (RMSE) are much better for data assimilation with EnKF than for the open loop simulations, both with and without parameter updating. The RMSE for the scenario where

Tab. 1 Perturbation of soil parameters and model forcing for CLM for each model layer (*N* is normal distribution and *U* is uniform distribution)

variables	Unit	Magnitude
Precipitation	Mm	×(1+0.1*N(0,1))
clay fraction	Percentage	+U[-10,10]
sand fraction	Percentage	+U[-10,10]
organic matter density	kg/m^3	+U[-15,15]
observation error	m^3/m^3	+N(0,0.02)

EnKF is also used to estimate parameters reduces, compared to the open loop simulations, 63% for layer 1, 80% for layer 2 and 86% for layer 3. Even slightly better results are obtained for the scenario where only states are updated. During the assimilation period, parameter values are allowed to change but changes in soil hydraulic parameters like saturated hydraulic conductivity were relatively small.

The same comparison was also made for the verification period. In this case, results for the open loop simulation and the simulations were only the state was updated in the assimilation period, were very similar, which indicates that the impact of improved initial conditions at the beginning of the verification period was limited. At 5cm depth the EnKF-runs with state updating only give similar errors as the open loop run, and at 20cm and 50cm depth RMSE are only slightly smaller (11% resp. 7% lower) than for the open loop simulation. The EnKF runs with parameter updating gave a much lower RMSE with errors 23%, 30% and 68% lower than for the open loop run, at 5cm, 20cm and 50cm depth respectively. The absolute errors were around 0.02 m^3/m^3 for the two lower layers and close to 0.05 m^3/m^3 for the upper layer. These results illustrate the feasibility of joint state and parameter estimation with land surface models.

However, for upscaling purposes it is not so trivial to use the estimated parameters. Whereas it can be argued that estimated ecosystem parameters for plant functional types can be used for other plant functional types, this is in general not the case for soil hydraulic and thermal parameters which vary spatially and can be quite different for the same soil type. Therefore remote sensing information, together with in situ data, plays a central role in upscaling of water balances at the basin scale.

Upscaling at basin scale with land surface temperature

A study for the Rur catchment illustrates the upscaling to the basin scale with help of remote sensing information and sequential data assimilation. In this case, land surface temperature measured by MODIS (Terra/Aqua) was used for assimilation. Land surface temperature is measured four times per day (twice during daytime and twice during night). The land surface temperature products are provided at high spatial resolution (1 km) and with an expected high accuracy (less than 1 K error - http://landval.gsfc.nasa.gov). Therefore, MODIS land surface temperature is a good data source for land data assimilation at catchment scale. The two-source formulation model (Kustas and Anderson, 2009) was used as the observation operator.

The sensitivity of sensible and latent heat fluxes to the CLM states/parameters has been evaluated in recent studies (Bonan et al., 2011; Hou et al., 2012; Schwinger et al., 2010). It was shown that the most sensitive CLM states/parameters were soil moisture, leaf area index, maximum rate of carboxylation (*Vcmax*), decay factor of subsurface runoff (*Fdrai*) and maximum drainage when the water table is at the surface (*Qdrai*). In this study, the data assimilation framework included the estimation of the mentioned three CLM-parameters. The leaf area index as provided by MODIS underestimates the true LAI by 0.66 (m^2/m^2) on average for all landcover types (http://landval.gsfc.nasa.gov). The parameter values of *Vcmax*, *Fdrai,* and *Qdrai* are tuned globally in CLM (Oleson et al., 2013). These values need to be calibrated for specific applications at the smaller catchment scale. Therefore, MODIS LST was assimilated in CLM to update these sensitive variables and parameters with joint state and parameter estimation.

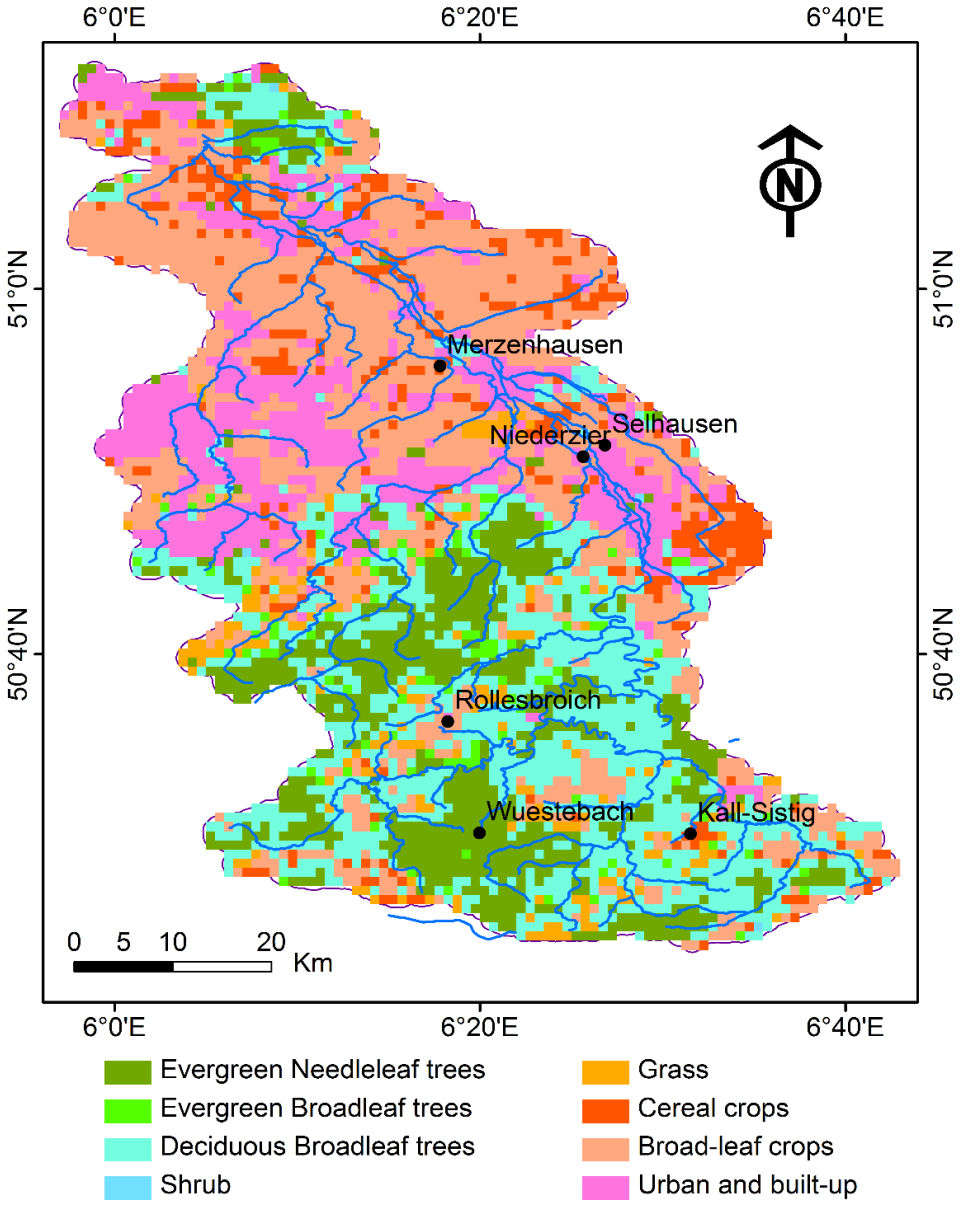


Figure 2. Landcover types for the Rur Catchment.

The data assimilation was now carried out for the complete Rur catchment (2454 km^2) located in the border region of Belgium–Netherlands–Germany, which includes the Rollesbroich site. See Figure 2. The annual precipitation and potential evapotranspiration for the northern part are in the range of [650 - 850 mm] and [580 - 600 mm], respectively; in the mountainous southern part, they are in the range of [850 -1300 mm] and [450 - 550 mm], respectively. The MODIS leaf area product (MCD15A3 - https://lpdaac.usgs.gov) and the soil properties of the Harmonized World Soil Database (FAO et al., 2012) were used as CLM input (Han et al., 2012; Han et al., 2014). The COSMO_DE reanalysis data from 2010-2012 were used as atmospheric forcing data (Baldauf et al., 2009).

The land surface temperature measurements to be assimilated in CLM were obtained from the daily MODIS land surface temperature product MOD11A1 in 2012

(https://lpdaac.usgs.gov). Only good quality data were used for assimilation, according to the data quality flag (average error <= 1 K), and 246 observations were assimilated in total.

The spinup of CLM was done for 10 years using the forcing data from 2010 and 2011 repeatedly. Data assimilation for the year 2012 was carried out for the following seven simulation/assimilation scenarios: (1) **Openloop** – CLM open loop simulation without data assimilation; (2) **LST_Vcmax** - update $Vcmax$ by assimilating LST; (3) **LST_Fdrai** - update $Fdrai$ by assimilating LST; (4) **LST_Qdrai** - update $Qdrai$ by assimilating LST; (5) **LST_LAI_SAI** - update LAI and SAI by assimilating LST; (6) **LST_Par_All** - update $Vcmax$, $Fdrai$, $Qdrai$, LAI and SAI by assimilating LST; (7) **LST_Feedback_Par_All** - update $Vcmax$, $Fdrai$, $Qdrai$, LAI and SAI and soil moisture by assimilating LST. The soil temperature profiles were also updated in all assimilation scenarios.

Fifty ensemble members of atmospheric forcing data, soil properties and leaf area index were generated with the random perturbation method. The parameter estimation of $Vcmax$, $Fdrai$, $Qdrai$, LAI and SAI was carried out with help of EnKF including state augmentation. The update of the soil moisture profile (scenario LST_Feedback_Par_All) was also done with the state augmentation approach. As $Vcmax$, $Fdrai$ and $Qdrai$ are hard coded parameter values in CLM, the CLM source code was modified to allow the changes of these parameters. The parameter values of $Vcmax$, $Fdrai$ and $Qdrai$ are [10.75 (µmol/m^2s) ~ 177.32 (µmol/m^2s)] (Walker et al., 2014), [0.1 ~ 5.0] and [10^{-6} ~ 10^{-1}] (Hou et al., 2012), respectively. In order to represent the uncertainties, multiplicative uniform distribution perturbations of U[0.75,1.25], log2(U[0.75,1.25]) and log10(U[0.75,1.25]) were used for $Vcmax$, $Fdrai$ and $Qdrai$, respectively. The perturbations were applied only when these parameters were calibrated in data assimilation.

As verification data sensible and latent heat fluxes measured by Eddy Covariance (EC) systems (Aubinet et al., 2012; Liu et al., 2011) at the stations of Kall-Sistig, Merzenhausen, Rollesbroich, Niederzier and Selhausen were used. In addition, soil moisture and soil temperature at three depths (5cm, 20cm and 50cm) averaged from 87 nodes of the wireless network SoilNet at Rollesbroich (Bogena et al., 2010) were also verification data.

Fig. 3 shows the RMSE values of latent heat flux for the open loop run and the different data assimilation scenarios. The RMSE values at the five EC-stations are summarized. In general, hardly any improvements could be achieved, except for the calibration of LAI and SAI. In this case the latent heat flux at four EC-stations (Kall-Sistig, Rollesbroich, Niederzier and Selhausen) was improved (for the scenarios of LST_LAI_SAI, LST_Par_All and LST_Feedback_Par_All). On the contrary, the RMSE at the Merzenhausen site increased slightly.

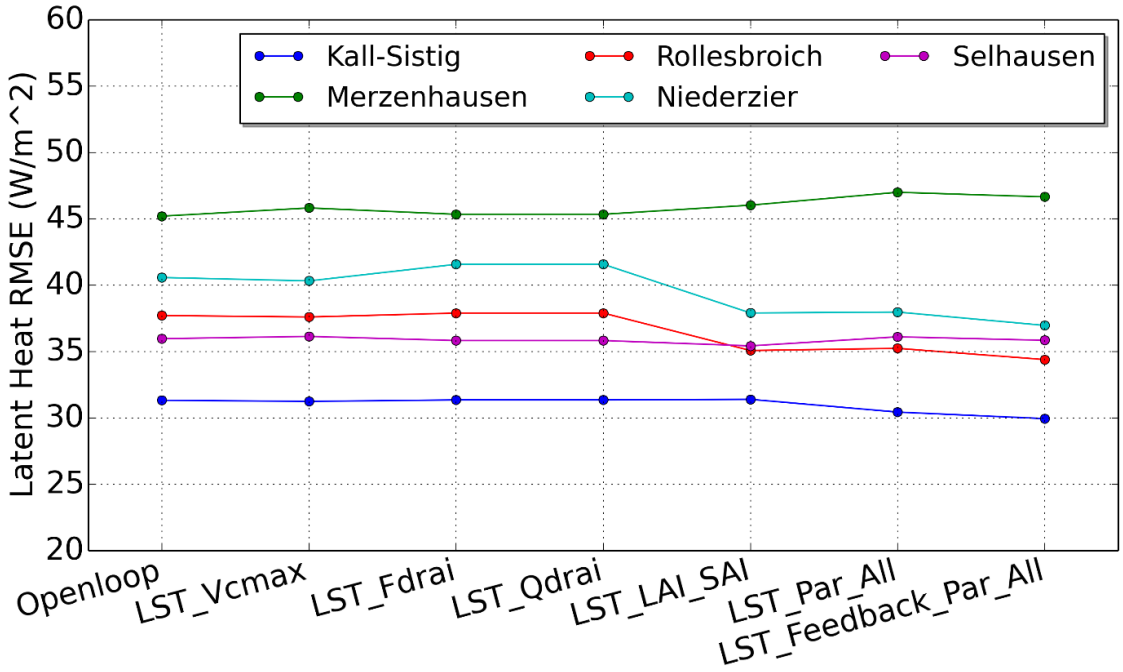


Figure 3. RMSE values of latent heat flux at 5 stations for the open loop run, and assimilation scenarios of joint LST and Vcmax estimation (LST_Vcmax), joint LST and Fdrai estimation (LST_Fdrai), joint LST and Qdrai estimation (LST_Qdrai), joint LST, LAI and SAI estimation (LST_LAI_SAI), joint LST, Vcmax, Fdrai, Qdrai, LAI and SAI estimation (LST_Par_All) and joint LST, soil moisture, Vcmax, Fdrai, Qdrai, LAI and SAI estimation (LST_Feedback_Par_All).

Results for the sensible heat flux characterization are illustrated in Fig. 4 for the same scenarios. In this case, a small reduction of RMSE values is achieved after the calibration of Vcmax, Fdrai and Qdrai parameters. However, the results for three sites (Merzenhausen, Rollesbroich and Selhausen) show a slightly worse performance for the scenarios of LST_LAI_SAI, LST_Par_All and LST_Feedback_Par_All compared with the scenarios of LST_Vcmax, LST_Fdrai and LST_Qdrai. The results are however not worse than for the open loop scenario. It seems that for sensible heat flux characterization, the calibration of Vcmax has the largest positive impact.

The RMSE values for the characterization of the soil moisture and soil temperature profile at the Rollesbroich site do not show a significant impact of data assimilation and parameter estimation. Only for the scenario LST_Feedback_Par_All a larger impact is found with an increase of RMSE for soil moisture content but a decrease of RMSE for soil temperature.

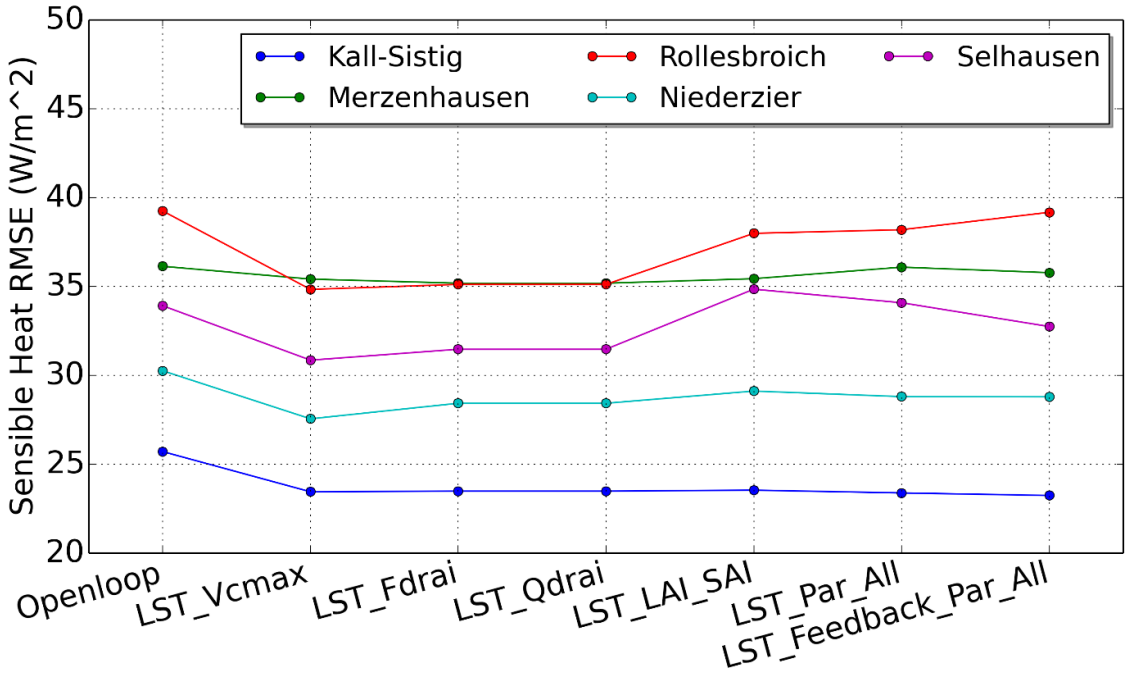


Figure 4. RMSE values of sensible heat flux at five EC-stations for different scenarios (see Fig. 3)

The updated LAI for five sites is shown in Fig. 5. The red lines show the LAI from the MODIS product, interpolated linearly by CLM. It is found that the underestimation of LAI by MODIS can be corrected by assimilating land surface temperature with the state augmentation method for these two sites. However, for the forest areas no improvement was found.

In summary, the feasibility of joint state-parameter estimation at the plot scale was demonstrated. For hydrological applications, upscaling is more difficult as parameters are spatially variable and vary also for the same soil type. Remote sensing plays a crucial role for upscaling of simulation results. An example was given for the Rur catchment, which showed a very slight improvement of simulated sensible and latent heat fluxes and leaf area indices.

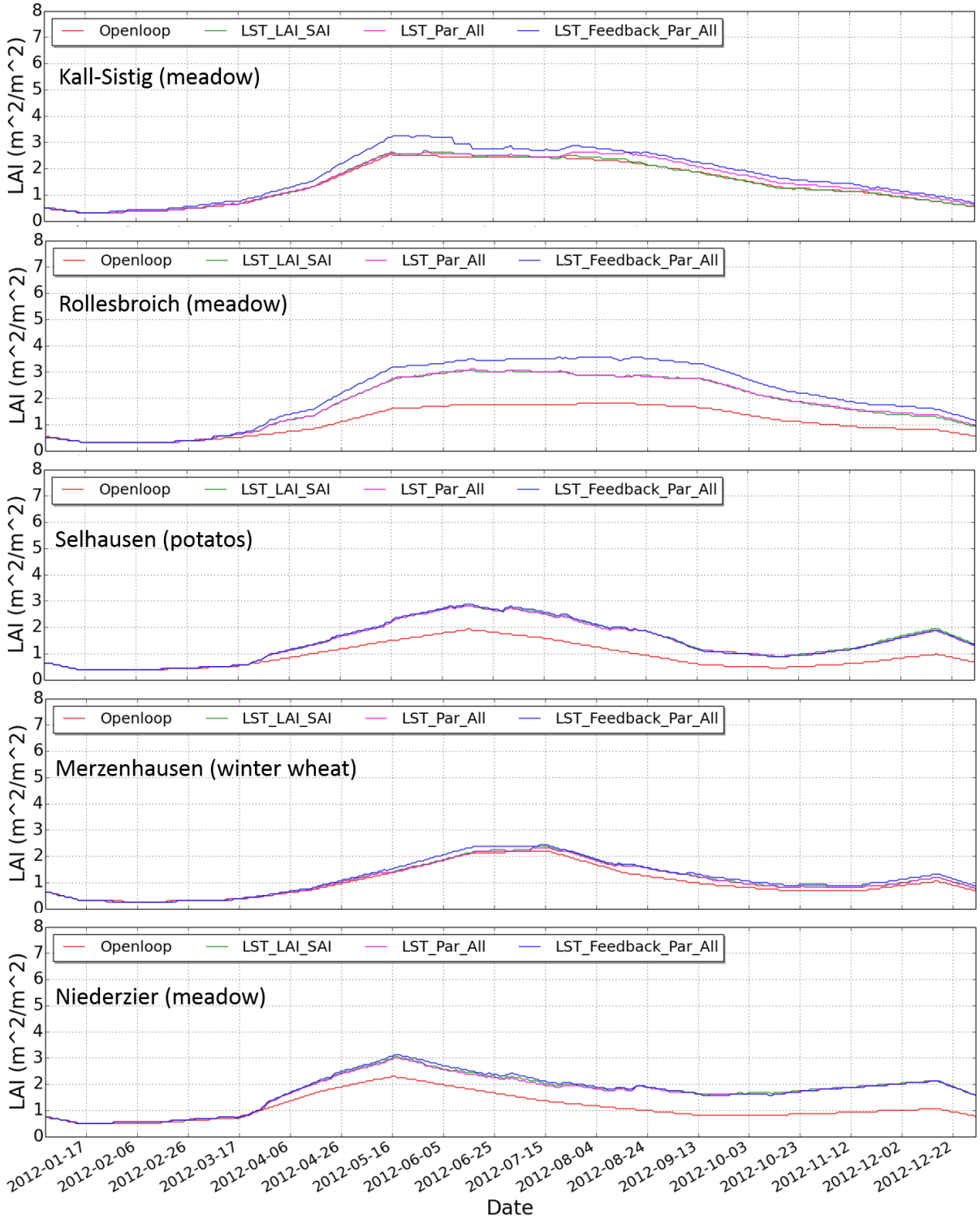


Figure 5. Updated leaf area index for four scenarios at the five stations.

4. Upscaling carbon fluxes and stocks at continental scale

4.1. How to optimize a process-based model: example with ORCHIDEE

The aim of this section is to describe the typical procedure followed when optimising the carbon stocks and fluxes of a Terrestrial Biosphere Model (TBM), of the type used in Earth System Models (ESMs), and to highlight typical issues that need to be considered in the process. Studies using the ORCHIDEE TBM (see http://labex.ipsl.fr/orchidee/), which is part of the IPSL ESM are shown as an example.

Note that the framework described below is built from the different examples and case studies summarized in the web-site http://orchidas.lsce.ipsl.fr/ (a tutorial is available at **http://orchidas.lsce.ipsl.fr/neon.php)**

4.1.1. Choosing observations

There are a growing number of data streams related to the C cycle that are available for optimising the parameters of a TBM. Typically they span a range of spatial and temporal scales (Figure 6) both at site level (or networks of sites), regionally (e.g. national inventories) and globally (via satellite measurements). Each different data stream can be used to constrain one specific component of the C cycle (e.g. stocks or fluxes) or an indirectly related process such as the hydrology.

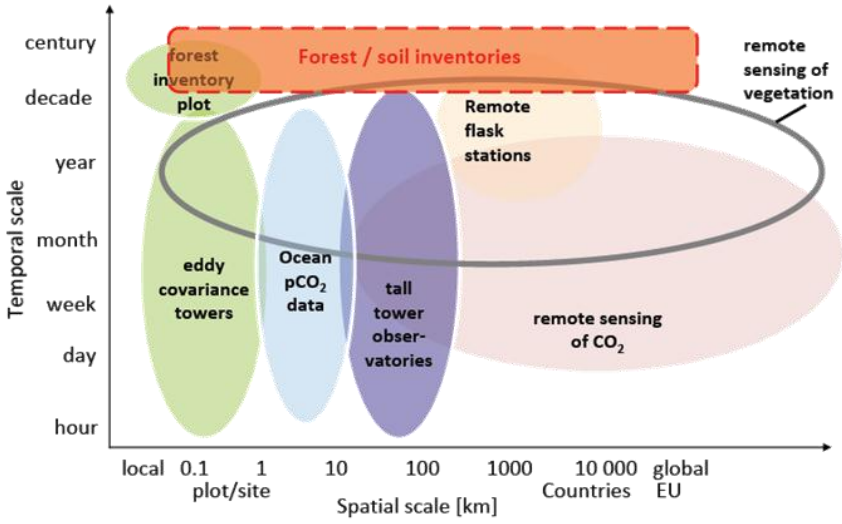


Figure 6: of the spatial and temporal constraint on the carbon cycle brought by different data streams

Below is a list and description of the commonly used data streams at different spatial scales.

Ground based site data

- Carbon and energy-related flux measurements (NEE, LE, SH): taken either using chamber-based instruments or eddy covariance meaurements at flux tower sites. The international FluxNet network gathers these data (as well as site-based information) for a range of sites/ecosystems across the globe

(Figure 7) and makes the data available on their website (http:// http://fluxnet.ornl.gov).

- Aboveground biomass measurements: determined from allometric relationships (e.g. related to tree "diameter at breast height") at certain sites on an annual basis, either as the total aboveground biomass or as the annual biomass increment. There is currently no database or network that collects these data into a common repository.
- Soil carbon stocks: same as for aboveground biomass measurements, except that measurements are not gathered as frequently due to the difficulty of measuring the soil C content of the soil.
- Soil moisture: volumetric soil moisture content is measured at various sites (many for validation of satellite-derived observations) using various methods and on various timescales depending on the method. Many data are gathered into the International Soil Moisture Network (https://ismn.geo.tuwien.ac.at).
- Atmospheric CO_2 data: Atmospheric CO2 concentration data are measured at various flask sampling sites globally (both on land and ocean) and the data from most stations are collected and provided through the GLOBALVIEW website (http://www.esrl.noaa.gov/gmd/ccgg/globalview/). These data integrate all components of the carbon cycle and provide valuable constraints at large spatial scale.

Satellite global observations

- Vegetation activity proxy: vegetation "greenness" indices such as the Normalised Difference Vegetation Index (NDVI)as well as "higher-order" more physically based products such as Leaf Area Index (LAI) and the fraction of absorbed Photosynthetically active radiation (fAPAR) data can be derived from satellite optical reflectance data, and give a measure of the productivity of the vegetation. These data can be principally used to give information on the phenology (leaf onset, senescence and growth rates) of the canopy (Weiss et al., 2007). Several instruments are flown on both ESA and NASA satellite missions.
- Soil moisture: satellite measurements in the microwave part of the electromagnetic spectrum used to derive soil moisture content of the vegetation and soil due the changing dielectric properties of both with increasing water content. Both active and passive instruments are flown on ESA and NASA satellites – with different algorithms used to derive the soil moisture (thus providing either volumentric or relative soil moisture observations). These data can be used to optimise parameters relating to the hydrology of TBMs, which in turn will impact the C cycle processes.
- Aboveground woody biomass (futur ESA mission, BIOMASS): Stand-level biomass can also be obtained from microwave satellite instruments. No such instrument currently exist, but ESA are planning to launch its BIOMASS satellite mission in ~2020. Note that few recent biomass estimates based on satellite data and in situ measurements have been derived for the Tropics (Saatchi et al., 2011) and for high latitudes (Thurner et al., 2014)
- Fire product (global burned areas): Burnt area can also be derived from optical reflectance data due to the abrupt change in the signal (both

temporally and spatially) once a fire has occurred. These products are mostly available from NASA instruments and more recently from an ESA initiative (https://geogra.uah.es/fire_cci/)

Regional or country based inventories

- Forest carbon stocks inventories: most countries report their forest carbon stocks to the FAO. However these estimates are still missing for some countries and remain highly uncertain.

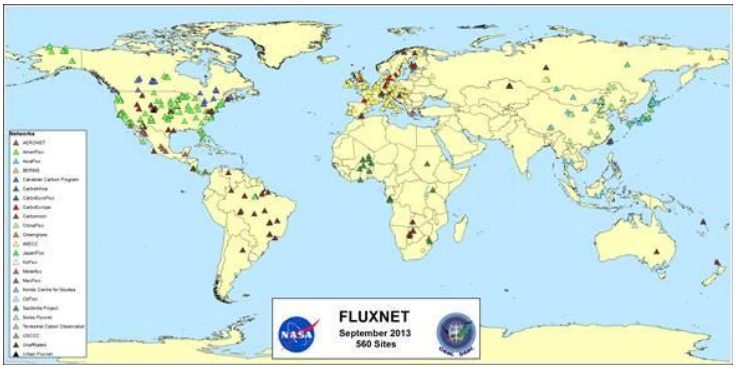


Figure 7: Location of the FluxNet sites (FluxNet network) with NEE, LE and LH measurements

Satellite data offer a clear advantage in that they cover the whole globe. However no physical property of the ecosystem can be measured directly – all products have to be derived from the raw radiance measured by the satellite instrument. Even to derive a measure of the surface reflectance an atmospheric correction must first be performed using a model to approximate the scattering of light by clouds, aerosols, greenhouse gases etc. To derive any information on physical properties a “radiative transfer model” is also required, which necessarily introduces uncertainty related to the adequacy of the model structure and uncertain model parameters. Ground-based measurements usually can measure the property more directly, though not always (see for example aboveground biomass data). However, it is impossible to gather measurements for all points across the globe. Although the number of sites which collect these data are increasing, and networks are being established in order to gather these data for ease of use, the sites are often located in developed countries where funding is more readily available, and as such not all ecosystems/biomes are well-represented. For example we can see in Figure 7 that there is clearly a bias in terms of the representativeness of ecosystems toward temperate and boreal ecosystem in the FluxNet network. In comparison there are few in Africa and Asia, and almost none in savannah ecosystems.

Further work needs to be done to extend existing measurement campaigns and to gather existing data for all data streams into a network. This will allow the data to be provided in a common format with useful metadata and to make the data easy to download and use.

In terms of optimising a model, it is crucial to select the right data stream for the scientific question that is being investigated. Inconsistencies may exist between the model and the data or between various data streams (examples are highlighted below). Furthermore, consideration should be given to the temporal and spatial resolution compared to the model, as otherwise unaccounted for biases will hamper the optimisation.

4.1.2. Selecting model parameters

We first need to select the model parameters that will be optimized given two constraints as practically, we can not optimize all model parameters as for example ORCHIDEE has over 200 parameters. Thus we need to select the most sensitive parameter for a given scientific objective. In this context, a sensitivity analysis is probably the first step that needs to be performed. It identifies what the uncertainty in the model outputs is attributable to in the model inputs (parameters in this case). This is achieved by running the model for many different random samples of the parameter space to determine their impact on the model state variables. It provides information on how much of the output variance is controlled by each parameter. The more complex, variance based SA methods (e.g. Sobol, FAST ETC) give a quantitative estimate of the percentage of the variance that each parameter is responsible for. Less complex "screening" methods provide qualitative information by ranking each parameter in terms of how great an impact they have on the model outputs. These methods are less computationally expensive. The Morris method is an example of a screening method that can efficiently determine the importance of the parameters and produces results comparable to the more complex methods.

The Morris method is based on determining incremental ratios, or "elementary effects", from which basic statistics are used to define the model sensitivity (Morris, 1970; Campolongo et al., 2007). The elementary effects are based on changing the value of one parameter at a time in sequence for many (*r*) trajectories which populate the parameter space. The difference in the model outputs is recorded and the mean of the absolute values of the changes (μ*) and standard deviation (σ) of the difference from all trajectories for each parameter is calculated once the sampling has finished. Thus this is a global search method which can be used to determine which parameters have a negligible impact on the model, those which have a linear and additive effect, and those which are non-linear and are involved with interactions with other factors (Campolongo et al., 2007).

Table 2 shows first the ensemble of parameters that are considered for the optimization of ORCHIDEE, with their typical prior value and a range of possible variation that has been determined based on expert knowledge of the code. A prior uncertainty is also given based on this expert knowledge

Parameter	Description	Prior value	Prior range	σ_{prior}
Photosynthesis				
V_{cmax}	Maximum carboxylation rate (μmol $m^{-2}s^{-1}$)	55	27-110	33.2
$G_{s,slope}$	Ball-Berry slope	9	3-15	4.8
T_{opt}	Optimal photosynthesis temperature (°C)	26	6-46	16
T_{min}	Minimal photosynthesis temperature (°C)	–2	(–7)–3	4
SLA	Specific leaf area (LAI per dry matter content, m^2g^{-1})	0.026	0.013-0.05	0.0148
LAI_{max}	Maximum LAI per PFT (m^2m^{-2})	5	3-7	1.6
$K_{lai,happy}$	LAI threshold to stop carbohydrate use	0.5	0.35-0.7	0.14
Phenology				
$K_{pheno,crit}$	Multiplicative factor for growing season start threshold	1	0.5-2	0.6
T_{senses}	Temperature threshold for senescence (°C)	12	2-22	8
$L_{agecrit}$	Average critical age for leaves (days)	180	80-280	80
Soil water availability				
Hum_{cste}	Root profile (m^{-1})	0.8	0.2-3	1.12
Dpu_{cste}	Total depth of soil water pool (m)	2	0.1-6	2.36
Respiration				

Q_{10}	Temperature dependence of heterotrophic respiration	1.99372	1-3	0.8
K_{soilC}	Multiplicative factor of initial carbon pools	1	0.1-2	0.76
$HR_{H,b}$	First-degree coefficient of the function for moisture control factor of heterotrophic respiration	2.4	2.1-2.7	0.24
$HR_{H,c}$	Offset of the function for moisture control factor of heterotrophic respiration	–0.29	(–0.59)–0.01	0.24
MR_a	Slope of the affine relationship between temperature and maintenance respiration	0.16	0.08-0.24	0.064
MR_b	Offset of the affine relationship between temperature and maintenance respiration	1	0.1-2	0.76
GR_{frac}	Fraction of biomass available for growth respiration	0.28	0.2-0.36	0.064
Energy balance				
$Z0_{overheight}$	Characteristic rugosity length (m)	0.0625	0.02-0.1	0.032
$K_{albedo,veg}$	Multiplying factor for surface albedo	1	0.8-1.2	0.16

Table 2 : List of considered parameters and prior value, range of variation and uncertainty.

As an example (figure 8) we applied the MORIS sensitivity analysis with respect to one model variable, the NEE at a given FluxNet site with Tropical Broadleaved Evergreen Trees: ecotone Bananal Island in Brasil (BR-Ban — http://fluxnet.ornl.gov/site/58).

The x-axis in Figure 8 shows the absolute mean change in NEE averaged across the different time periods with respect to each parameter, and the y-axis shows the standard deviation and thus the interactions between parameters. In order to determine a simple ranking of the most important parameter, we consider the order according to the mean changes (i.e. the x-axis values). Figure 8 demonstrates that the relative importance of the parameters is different depending on the time period examined. This is obvious when considering the model physics – at nighttime for example, there is no C uptake, only respiration, and thus the parameters to which respiration processes are sensitive have the highest value of mean change. The results highlight the necessity therefore for considering which timescale we most want to focus on in the optimisation.

Figure 8 shows that for most time periods, the soil depth (Dpu_cste) is an important parameter. Other parameters that we may also expect to be important in the daytime, such as the optimium C assimilation efficiency (Vcmax_opt), or the specific leaf area (SLA) and a parameter controlling the fraction of maintenance respiration (Maint_resp_slope_c) are also dominant across different timescales. Following that, other parameters that are related to the soil moisture status (such as the rooting depth profile – Humcste) are also important. This is an interesting result as it shows that processes that are not directly involved in the C cycle such as the hydrology, which indirectly limit the rates of C uptake and decomposition processes, also factor highly. This suggests that we should consider optimising the hydrology of the model, using soil moisture or related observations, alongside the carbon fluxes might be of considerable benefit.

One further point to note is that in addition to the time period considered the sensitivity analysis also depends on the range that is set for each parameter. If there is a large part of the range of parameter values to which the model variable is not sensitive, this will dilute the importance of this parameter in the sensitivity analysis. Equally, if the model is highly sensitive to regions close to the upper or lower boundary of the parameter range, the overall importance of that parameter may be overestimated, as it is only in extreme situations that the model will enter into that part of the range. One example of this is parameter values, which would give extremely low soil moisture values. Thus the choice of the parameter range that is included in the sensitivity analysis should be considered carefully before performing the sensitivity analysis.

The final point to note when conducting a sensitivity analysis is that the results will be different between different sites within the same PFT, mostly due to different climatological forcing. The model may be more or less sensitive to some parameters depending on the environmental limitations imposed by the climate. Therefore it is imperative to do the sensitivity analysis at a range of sites.

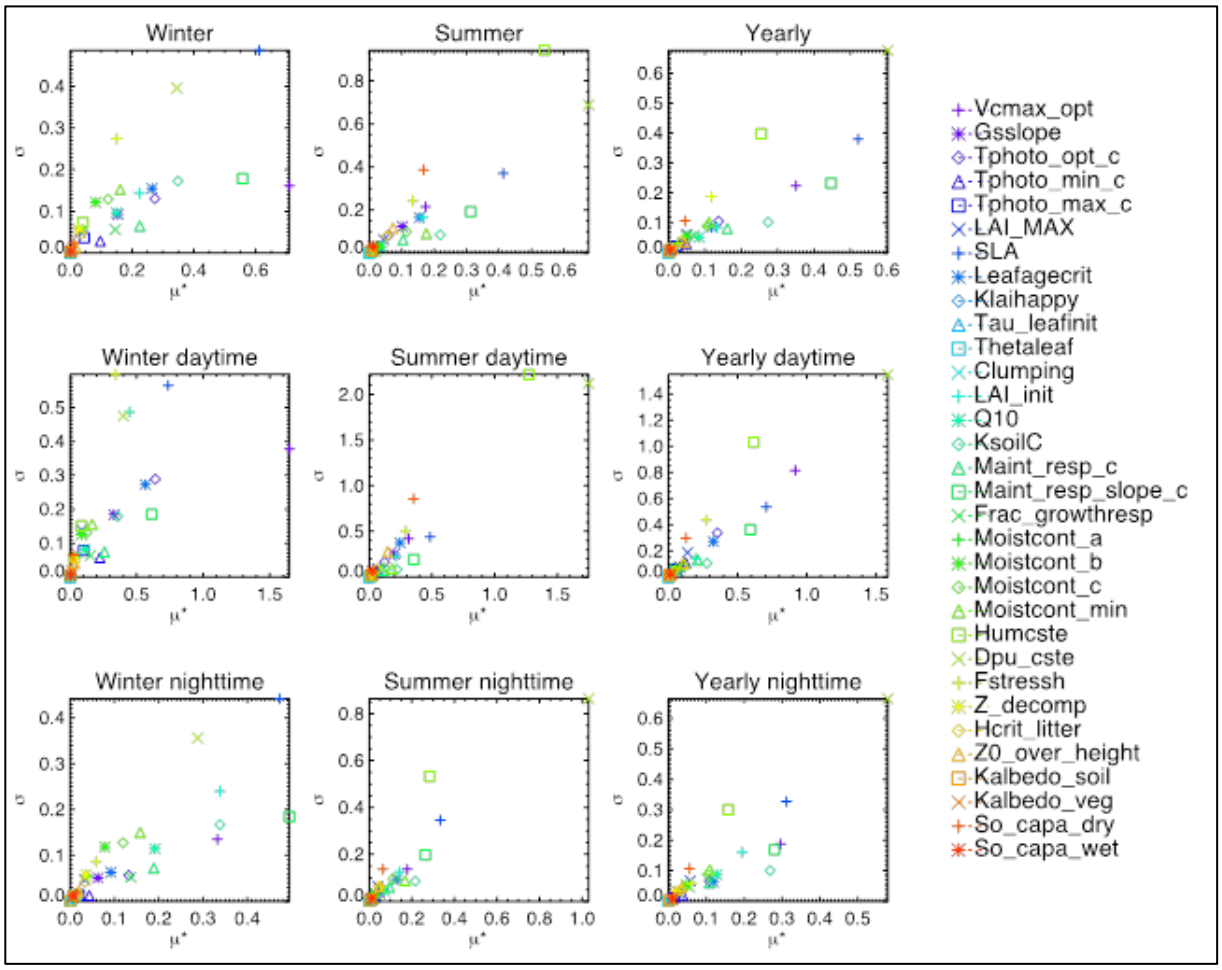


Figure 8: Result of the MORIS sensitivity analysis at a fluxnet site for the NEE; Result are displayed for different type of model NEE averaging and period: winter, summer, yearly and then separating between daytime and night time means.

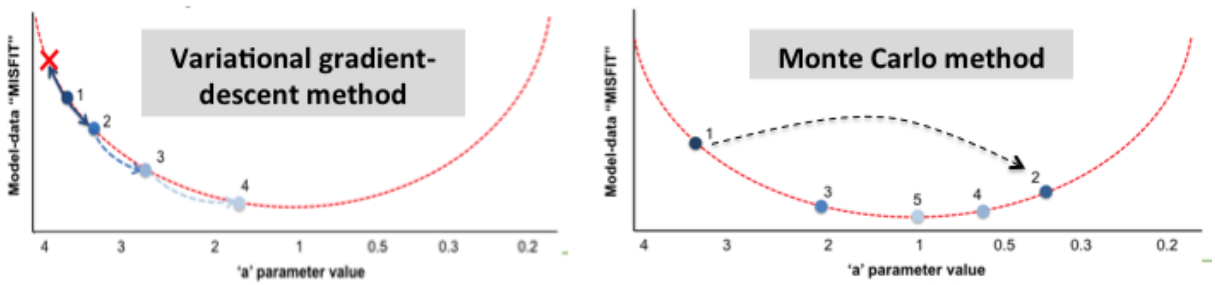
4.1.3. Finding the optimal parameters and their uncertainties

As introduced in section 2.2.1, the aim of the assimilation procedure is to minimize a misfit function that measures the mismatch between i) the model outputs and the various data streams and ii) *a priori* knowledge of the parameter values, taking into account uncertainty of both components in a statistically robust framework. Assuming that the errors associated to the parameters, the observations and the model outputs are Gaussian, the optimal parameter set corresponds to the minimum of a misfit function, J(x) (Tarantola, 1987, Eq. 1):

$$J(x) = \frac{1}{2}[(H(x) - y).\mathbf{R}^{-1}.(H(x) - y)^T + (x - x_b).\mathbf{P}_b^{-1}(x - x_b)^T] \quad (1)$$

where *H* represents the model that operates on the parameter vector, *x*, providing the model state variable (H(x)) that corresponds to the observation, *y*. **R** and $\mathbf{P_b}$ represent the error variance/covariance matrix associated to the observations and the parameters.

Finding the minimum of J(x) can be illustrated geometrically (see figure 9). This figure shows two different classes of different methods which can be used to find the minimum of the cost function. The gradient-descent method uses the gradient of the cost function to determine in which direction to move in the parameter space. An iterative algorithm will eventually find the minimum of the cost function ('a' parameter value = 1). The Monte Carlo, or any "global search" method, will randomly sample the parameter space, with no information on the shape of the cost function, and will compare the misfit between the observations and the model for the latest parameter value compared to the previous. The difference between these two algorithms are highlighted in the top panel of Figure 9. Depending on the degree of non linearity of the model, the shape of the cost function can be very different and thus require different algorithm to find optimally the global minimum (bottom panel in Figure 9).



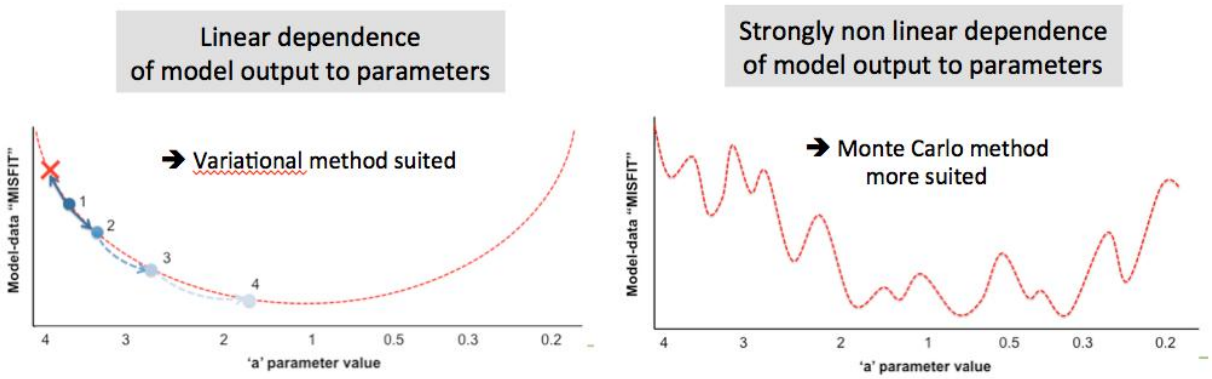


Figure 9: Top panel: The difference between the variational gradient-descent method and the Monte Carlo "global search" method for finding the minimum of a simple 1-dimensional cost function for a linear problem. Bottom panel: which method is appropriate for a linear vs non-linear cost function.

For a global TBM such as ORCHIDEE it is not always practical to perform an optimisation using a global search method as it is too computationally expensive when many sites/points are included in the optimisation. However as the bottom panel in Figure 9 highlights, the gradient-descent method may not be able to find the global minimum when the cost function is highly non-linear, as is likely the case for a complex model.

In a recent study by Santaren et al. (2014) the ability of a variational gradient-descent algorithm was compared to the Genetic Algorithm global search method at one site. The experiment outline was as follows:

- Pseudo-Data with randomly perturbed parameters, within their allowed range of variation (either 50% or 100% - see figure below).
- 10 experiments starting with different prior parameter values (randomly derived from the allowed range of variation), using 2 different methods:
    - Variational scheme: Iterative minimisation using the gradient of J at each iteration (obtained with the Tangent linear model); minimization with BFGS algorithm using 40 iterations
    - Monte Carlo "global search" method: Genetic Algorithm (GA)
    - Comparison of 1 year of data with 4 years

Figure 10 shows the results of these experiments. Overall the Genetic Algorithm performs better for all random first guess parameter values (low values and narrow spread in the prior cost function reduction) than the Variational (BFGS) approach. This is true when the random first guesses are taken from 50% or 100% of the parameter range or whether 1 or 4 years of data are used. At least for this site therefore, it would be better to use the GA, as the cost function is highly non-linear. This is possible at one site, but with a multi-site optimisation the GA is not practical from a computational cost point of view. However, it is hypothesised that if many sites were included in the same cost function, the variational approach would perform better than at a single site, as the cost function overall would be smoother. This is the subject of an on-going study using the ORCHIDEE model.

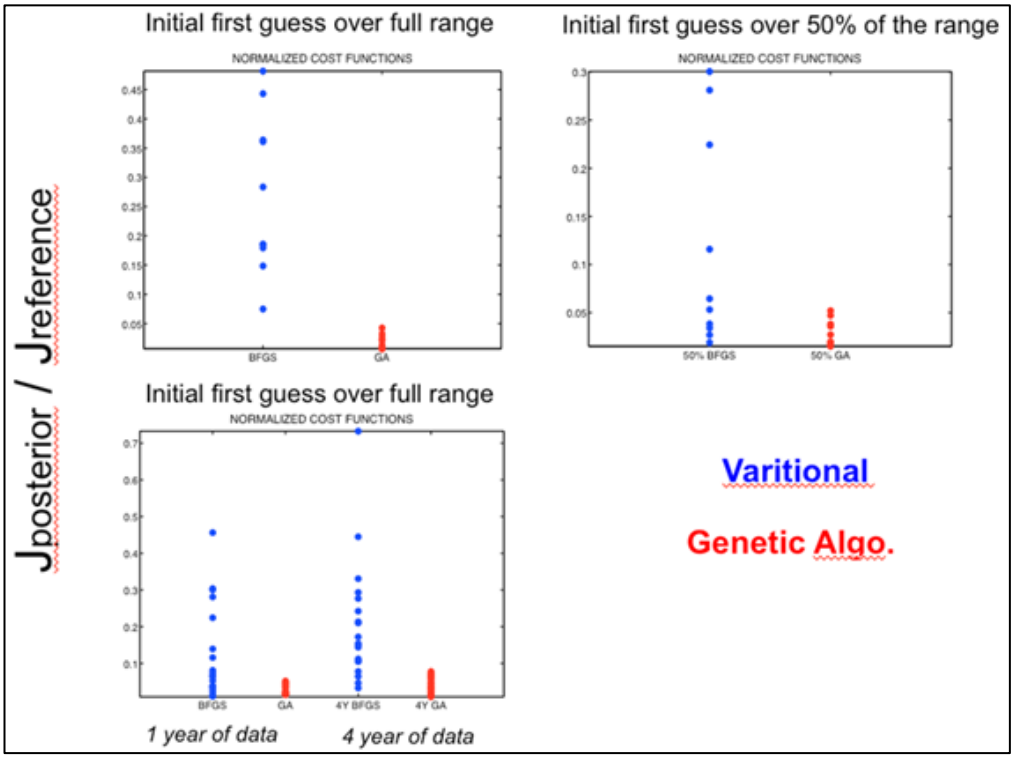


Figure 10: Cost function reductions for 10 twin experiments that were performed with the BFGS algorithm (BFGS) and with the Genetic Algorithm (GA). Cost functions were normalized by the value of the cost function representing the mismatch between the synthetic data and the model outputs computed with the standard parameters of ORCHIDEE (ratio of the posterior cost function (after 40 iterations) with respect to the cost function with ORCHIDEE standard parameters)

4.1.4. The need to assimilate multiple sites

In order to perform global simulations we need just one value per parameter per PFT, not the range of values that we would obtain if we optimised at each site individually. However, if we include all the sites in a multi-site (MS) optimisation, will we retrieve a generic parameter vector that can achieve the same reduction in model-data misfit at each site as if we only used the observations from that site? This question was investigated in a study by Kuppel et al. (2012, 2014) who used the variational (gradient-descent) method to optimize different PFTs of the ORCHIDEE model using FluxNet data from multiple sites and compared it to the case where we just optimize using the data at each individual site, in order to examine if we can obtain the same fit to the observations (Figure 11). Both NEE and LE daily fluxes were assimilated both at each site (single-site — SS) and for *all sites* for the same PFT (multi-site — MS — optimization). Many studies have used FluxNet data to optimize ecosystem/land surface models in the past, but none have used all the sites together in the same optimization for such a range of PFTs. A MS optimization is important, because for global simulations we need to have one parameter vector per PFT. If we just optimize at each site, we may find a range of posterior parameter values for each parameter for the same PFT. In order to do global simulations therefore, should the average be taken? Or is it better to include all sites in the same optimization? The aim of this study is therefore to see if the MS optimization performs as well as the SS optimization and/or the average of the SS optimisations.

The number of years assimilated is different for each site. The parameters that were optimized were related to the photosynthesis, respiration, energy balance and phenology. It

was assumed that there were no correlations between parameters. The errors on the observations were given as the prior RMSE between the model and the observation, and were assumed to be independent (i.e. no correlation).

Figure 11 shows results for NEE and LE for two of the twelve sites included in the optimization for the temperate broadleaved deciduous PFT. These results are described in Kuppel et al., (2012). In a further study (Kuppel et al., 2014), between 2 and 24 sites have been assimilated per PFT from stations across the globe depending on the data availability. The observations collected at each site are shown in black. The green curve shows the prior model simulation, the red the posterior model simulation for just the SS optimization (i.e. when only the data from that site were included in the optimization) and the blue shows the MS optimization (I.e. using the posterior parameter values when all 12 sites were included in the optimization).

Both the SS and MS optimization results in an improvement in the model-data misfit. In more extreme situations such as the summer of 2006 at Harvard forest, neither the prior or optimized model is able to fit the observations; therefore there is likely further need for modification to the model physics (e.g. possibly enhance uptake of C due to slight changes in climate or biotic effect not properly accounted for in the model).

The MS optimization does nearly as well as the SS in most cases, except for the LE at Fontainebleu. When the single site performs better, it indicates that potential improvement could be made if there the Plant Function Type parameter set would be further divided into different species-based or ecosystem-based parameter sets.

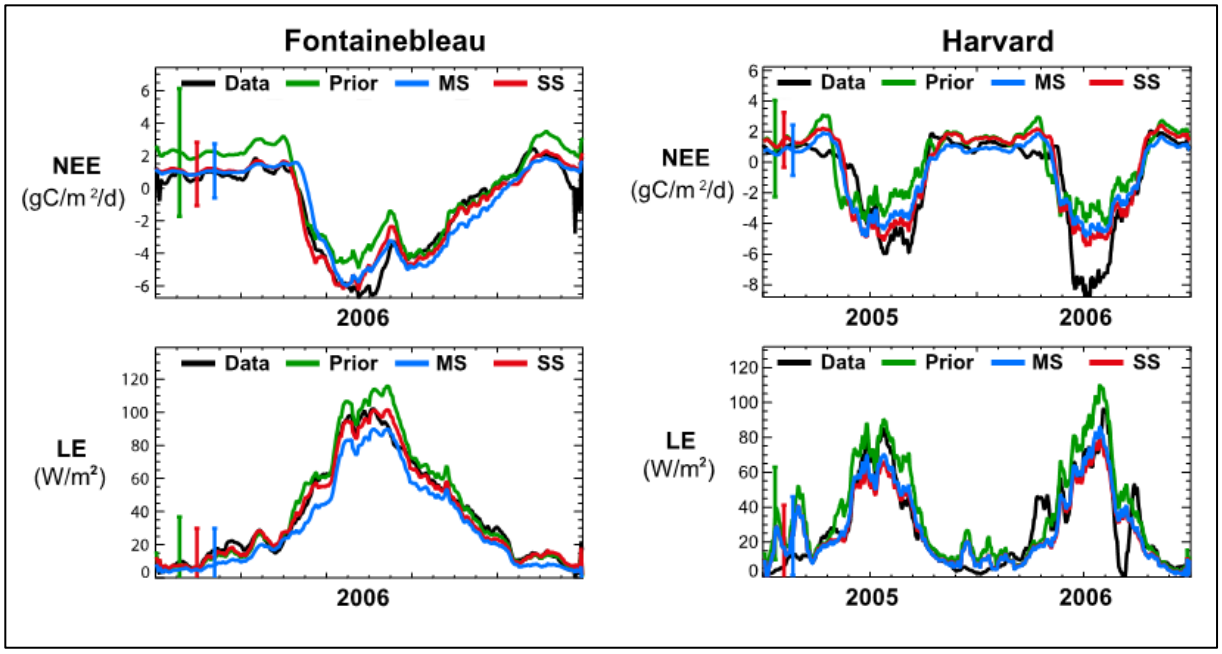


Figure 11: NEE and LE for two temperate broadleaved deciduous sites showing prior and SS and MS posterior model simulations compared to the observations.

Figure 12 shows the comparison of RMSD between the model and the observations for NEE and LE at all 12 sites using posterior parameters derived at each site with those derived at other sites, an average of all the SS values, and the MS posterior, with the prior. The aim is to see how well the MS posterior parameter values compare to the SS, and to the case where values obtained at another site are used. It is encouraging that the MS does nearly as well as the SS at most sites, and better or the same as the average of the SS values. In

addition, using posterior parameter values from SS optimisations at other sites (yellow bars) generally result in a worse RMSD than the SS or MS parameters, and can even degrade the fit compared to the prior (green bar). Each site has a different species and there are quite large "functional" differences between species of the same Plant Functional Type (like between Beech and Oak for the Deciduous temperate broadleaf forest). The sites also have different histories, characteristics and climate, which results in certain parameters becoming more or less important, and/or easy to find in the optimization.

These results strongly demonstrate that a MS optimisation is valuable and can find a unique, generic parameter set that largely results in a similar improvement in the model-data misfit. It also highlights that using one site to optimise a model is not sufficient. Of course, the MS posterior parameter will also be more robust as more sites are included.

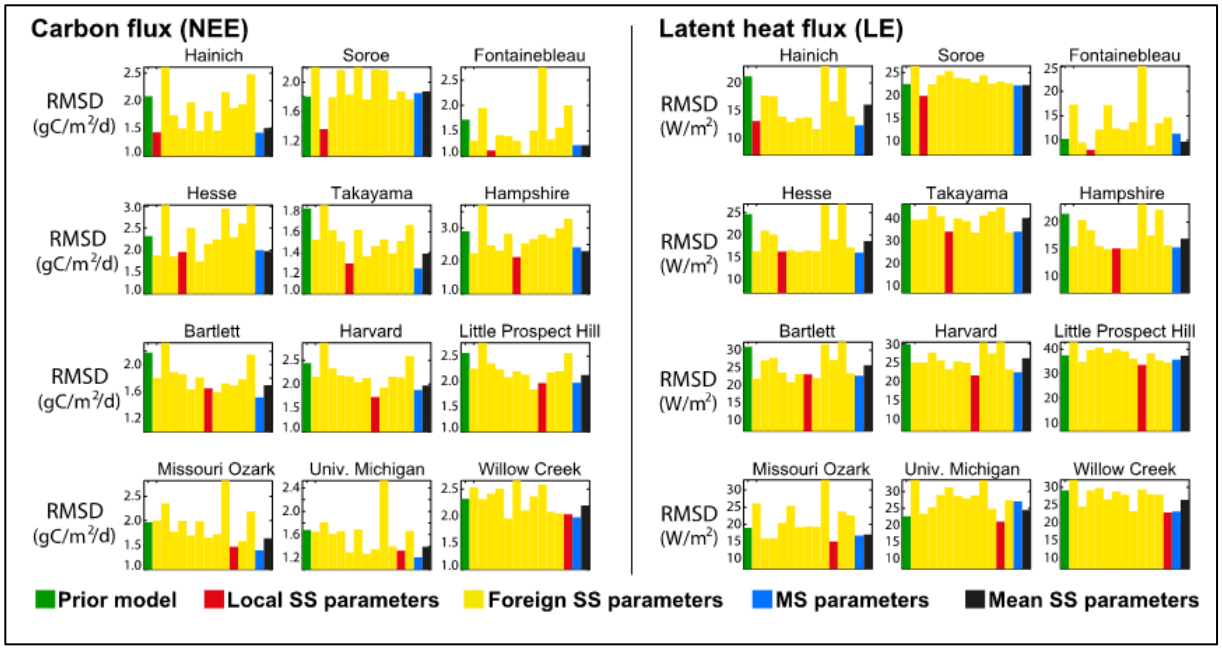


Figure 12: The RMSD between the model and the observations at each temperate broadleaved deciduous site with the prior parameters (green), the parameters from the SS optimisation at each site (red), the parameters from the MS optimisation with all sites (blue) and the parameters from every other site (yellow bars) for NEE and LE. The black bar also shows the RMSD when a mean of the SS parameters is used (rather than the MS).

The range of parameter values derived at each site compared to the prior and MS optimisation are shown in Figure 13 for each parameter. The MS posterior parameter value (black) is not always the same as the average of the values found at each site in an individual optimisation (colours), further suggesting that the MS optimisation can find a unique parameter vector.

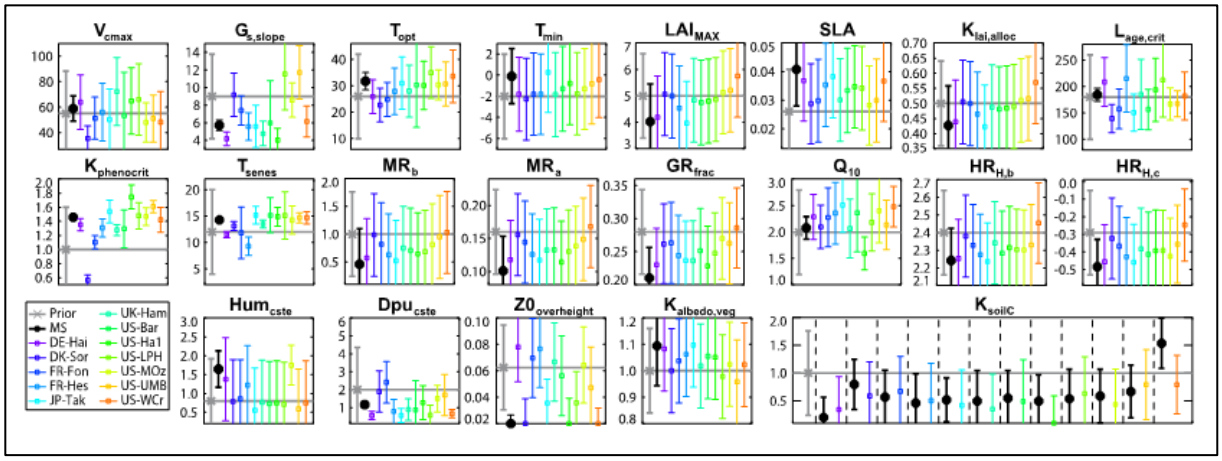


Figure 13: Prior (grey), MS (black) and SS (coloured) posterior values for each parameter included in the optimisation.

4.1.5. How to combine multiple data streams?

The use of multiple data streams is crucial to optimize model parameter relevant of different temporal and spatial scales. However technical complexity arises when dealing with different observations type that might complicate the overall optimization process. We detail below two approaches that have been established at LSCE with the ORCHIDEE model and provide an illustration of typical applications in both cases.

Simultaneous optimization of several data streams:

Measurements of C stocks, in addition to net C flux and latent heat flux measured at flux tower sites, are useful for optimising parameters related to slower processes in TBM, such as allocation and turnover of C. In one study with the ORCHIDEE model, observations of the total aboveground biomass, as well as the annual increment, were used together with flux measurements for such a purpose. The annual increment was able to offer an additional constraint compared to just the fluxes (Figure 14). The total aboveground biomass was used in a 2nd step to optimise the turnover rate of the soil C decomposition. However the total and incremental aboveground biomass observations are incompatible with each other; whereas the increment is always positive (Figure 14), the total biomass was adjusted for management (thinning) and natural influences (such as storms) that resulted in an overall decrease in biomass of the forest in some years. This resulted in an inconsistency in the model with respect to both data streams. In addition, although the fit of the model improved with respect to the total aboveground biomass, the posterior turnover rate was unrealistically high, due to the fact that the abovementioned processes that resulted in a decrease in the forest stand over time are not accounted for in the model.

This example highlights some of the issues faced when using different types of observations: the consistency between the model and all sources of data, which may be the result of incorrect processes in the model and/or incompatibility of the different data streams.

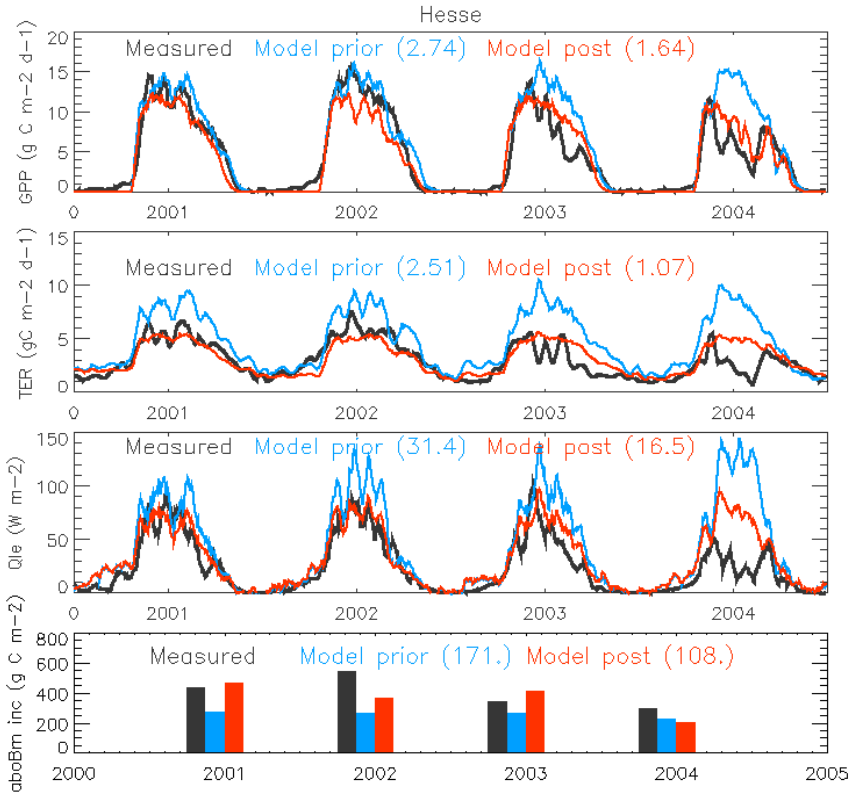


Figure 14: The micrometeorological fluxes (GPP, TER and Qle) and annual aboveground biomass increments at Hesse in 2001-2004. The 15-day running averages of the flux daily values are shown. The measurements are shown in black, a priori simulation result as green and a posteriori in red. The RMSE value between simulation and observation is shown in parenthesis.

However, when applying the optimization "globally" with multiple sites or global data or integrated constraint such as atmospheric CO2 concentrations, several difficulties appear. First, there is technical difficulties with the assimilation of all data streams at once, especially given the combination of in-situ local observations and global observations, which requests to run ORCHIDEE at point scale and globally. The characterization of the information content of each data stream separately indicates possible strong, complicated, non-linear interactions between the impacts of all parameters on the different outputs. This may hamper efficient convergence of the optimization algorithm to the optimal solution. Given this complexity, an efficient adjoint code of the ORCHIDEE model (i.e. optimized in terms of memory load and speed) would be needed in order to be able to perform a high number of iterations to properly assimilate all data streams in one step. However, only the tangent linear code of ORCHIDEE is yet available. Moreover, the join optimization requires more computing resources at a given time (memory load and processors), which increases the risk of failures during the iterative process. Secondly, the joint assimilation can hide the potential of each individual data stream (when mixed with the other sources of information), and thus requires strong expertise and knowledge of the model strengths and weaknesses in order to appropriately weight the contribution of each data stream.

Step-wise optimization of several data streams:

Given the difficulties mentioned above with a simultaneous approach, we propose to use a stepwise approach where all data streams are assimilated sequentially (i.e. one after the others). The information retrieved on the parameters at a given step (i.e., mean value and estimated error covariance matrix) is propagated to the next step and used as prior

information. Figure 15 illustrates such concept and the three steps that have been so far implemented for the ORCHIDEE carbon cycle data assimilation system:

Step 1 – MODIS-NDVI: Assimilation of a remotely sensed product of vegetation greenness (NDVI) from MODIS instrument. Only four parameters of ORCHIDEE for each Plant Functional Type (PFT), related to the seasonal cycle of the vegetation (phenology), are optimized.

Step 2 – FluxNet data: Assimilation of in-situ eddy-covariance net CO_2 (NEE) and latent heat (LE) flux measurements, from the FluxNet database (Baldochi, 2008). Daily fluxes for 78 sites, including several years per sites are used to optimize the most sensitive model parameters (see Table 2), controlling the fast carbon and water processes (photosynthesis, respiration, carbon allocation). ORCHIDEE is driven with the half-hourly meteorological data measured at each corresponding flux tower.

Step 3 – Atmospheric CO_2: Assimilation of atmospheric CO_2 measurements from 53 surface stations. We use the LMDz atmospheric transport model to assimilate the atmospheric CO2 concentration gradients in order to provide a final correction of the ORCHIDEE parameters. The objective is to provide an overall constraint to the land surface fluxes in order to match large-scale atmospheric constraints (i.e., global CO_2 growth rate, mean seasonal cycle and its latitudinal variation, as well as large scale north-south and east-west annual mean spatial gradients). The parameter vector defined for step2 is further optimized, using the posterior information from step2 optimization as prior information for step3, and expanded to include 30 parameters scaling the initial soil carbon pools for large “eco-regions”. The air-sea fluxes as well as fossil fuel and biomass burning emissions are also accounted for in this final step, in order to close the global carbon budget.

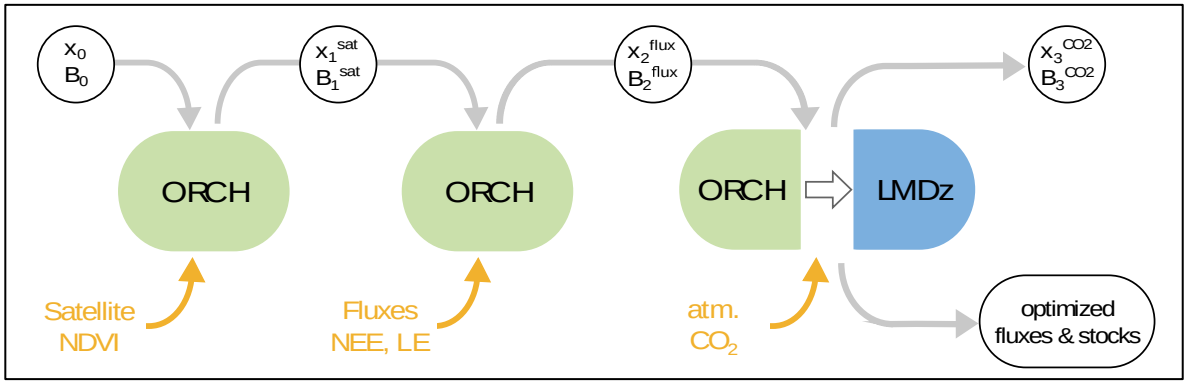


Figure 15. Illustration of the step-wise data assimilation approach used for the assimilation of multiple data streams in the ORCHIDEE-CCDAS.

The reasons for adopting a stepwise approach instead of a statistically more coherent simultaneous optimization follow from the limitations/difficulties discussed above. Additionally it is easier and straightforward to expose restricted sets of parameters to each observation type in the stepwise approach, in order to insure that each data stream constrains the parts of the model to which it has the most sensitivity.

If the system were completely linear (i.e. linear dependence of the flux with respect to the ORCHIDEE model parameters), the step-wise approach would be strictly equivalent to the simultaneous optimization. Unlike the atmospheric component, the land surface model is non-linear. Hence the step-wise assimilation of the various data streams may result in a suboptimal estimate of the final posterior parameter values due to inconsistencies between data streams, which result in a degradation of the fit to observations used in earlier steps of

the assimilation, as suggested in Kaminski et al. (2012). Moreover, simplifications have also been made in the stepwise approach for computational efficiency and practical issues. For instance, in step2 the optimization is performed for each PFT separately, although some parameters are PFT independent (like Q10, see Table 2). Hence after step2 we averaged the values estimated for each PFT in order to derive a global value to be used in step3. All these simplifications require to check after the last step that the final optimized set of parameters produces close to optimal results with significant model improvement for the all data streams that are considered (i.e., only slight degradation of the fit to the data compared to the optimization done specifically at each step). The step-wise system that has been built at LSCE verifies these constraint, at least with the 3 data streams that are considered (Peylin et al., in preparation).

4.2. Using an optimized model to upscale carbon fluxes and stocks

Upscale quantities and their uncertainties

The optimized model obtained after the assimilation of in-situ and/or satellite observations can then be run for any spatial domain (and at any spatial resolution) depending on the availability of meteorological forcing and land-cover description. We do not specifically discuss in this report the up-scaled carbon fluxes and stocks as they will be described in more details in the following deliverable D10.3.

One crucial point to mention is that the optimization process not only provides optimized parameter but also their associated uncertainty. We can thus propagate the uncertainty on the parameters (prior and posterior estimates) to the state variables (fluxes or stocks), using the linearity assumption.

The posterior parameter error covariance matrix, **Pa**, can be approximated to the second derivative (Hessian) of the cost function, using the linearity assumption at the minimum of *J(x)*. It can be derived with the Jacobian of the model, $\mathbf{H}_{\infty}$, following Tarantola (1987):

$$\mathbf{P}_{post} = [\mathbf{H}_{¥}^{T}\mathbf{R}^{-1}\mathbf{H}_{¥} + \mathbf{P}_{b}^{-1}]^{-1}$$

The posterior parameter error covariance can then be propagated into the model state variables space, $\mathbf{R}_{post}$, given the following matrix product:

$$\mathbf{R}_{post} = \mathbf{H}.\mathbf{P}_{post}.\mathbf{H}^{T}$$

The square root of the diagonal elements of $\mathbf{R}_{post}$ corresponds to the standard deviation, σ, on carbon fluxes/stocks of each grid cell. In order to appraise the knowledge improvement brought by the assimilation, the error reduction is usually discussed: $1 - (\sigma_{post} / \sigma_{prior})$.

Evaluation of the optimized model

Once the parameters have been optimized, it is crucial to validate the "optimized model" with independent data, i.e. data that have not been used in the assimilation process. Indeed, you could always have optimized some parameters to compensate for structural model errors which may degrade your overall model skill. It is thus important to keep some observations for an evaluation step.

As an example, figure 16 provides an example of global carbon flux evaluation after having optimized the parameter from the ORCHIDEE model using in-situ eddy covariance flux measurements (NEE and LE; Kuppel et al. 2012). The global net ecosystem exchanges have been transported with the LMDZ transport model before the optimization (prior curve) and after the optimization (Multi-site curve) using also the contribution from air-sea fluxes and fossil fuel emissions. The resulting atmospheric mean seasonal CO2 cycle at several stations is evaluated against the observations from the NOAA network. This example illustrates the gain obtained at the Alert station (North Canada, (high latitude)) and the degradation at Mauna Loa station (Hawai island) of the fit to CO2 concentrations. For Mauna Loa, it clearly indicates different possible flaws: using fluxnet observations may not be sufficient to optimize the net ecosystem exchange simulated by ORCHIDEE for the temperate northern hemisphere; model structural errors (in ORCHIDEE or possibly from the transport model) may hamper an efficient assimilation of the fluxnet observations, or both. Although it is difficult to assess the importance of model structural errors, the evaluation against independent data streams always brings valuable information.

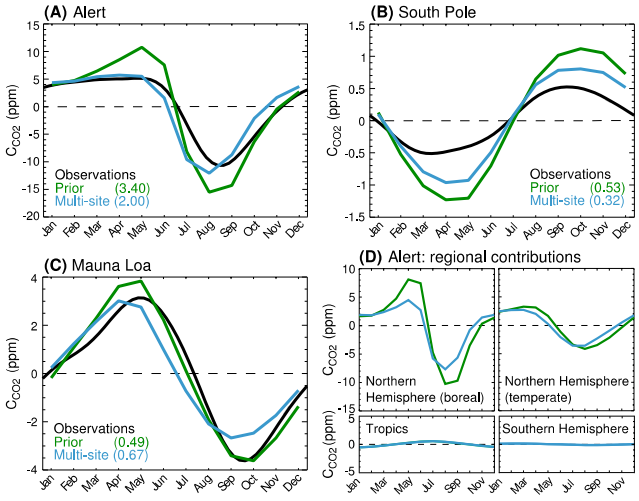


Figure 4.12: De-trended mean seasonal cycle of the atmospheric CO_2 concentrations at (a) Alert, (b) South Pole and (c) Mauna Loa locations during the 1989–2009 period: the optimization-independent concentration records (black) are compared to simulations where the biospheric contribution is calculated using the ORCHIDEE model with the default (green) and multisite optimized (blue) parameterizations, and the model–data RMSD given between brackets. (d) Regional contributions to the mean seasonal cycle simulated at Alert (Kuppel et al., 2014).

5. Conclusion / Next steps

This reports describes different frameworks to upscale biogeochemical fluxes at large spatial scale, using in-situ observations. Rather than a comprehensive literature survey of all possible up-scaling approaches, we described the principles of two methods, as well as their practical implementation, based on the use of a process-based ecosystem model (LSM or

ORCHIDEE). Two case studies were detailed: one for water fluxes for a small catchment and one for carbon fluxes at continental scale.

Experimentalists or modelers can thus use this report as a tutorial that will guide them into Data Assimilation (DA) approaches in order to understand the strengths and weaknesses of these frameworks. These can be resumed as:

- Model state optimization is a powerful approach to combine information embedded into current process-based models with observations of the state variables; however it does not provide improved predictions and/or analysis skills. It has been extensively used for the water cycle.
- Model parameter optimization provides alternatively new skills to upscale ecosystem properties (fluxes and stocks of water, carbon and energy). However, this technique deeply relies on the model structure and whether it captures the main processes controlling the transfer of carbon and water in the soil-plant-atmosphere continuum. Any structural model error will directly impact the optimized parameters and the resulting model skills. For the water cycle, the heterogeneity of soil hydraulic properties may also hamper any upscaling framework.
- Optimizing model state variables or parameters requires complex algorithms that need to deal with non linearity between model parameters and model state variables. For a proper use of these algorithms, the users need to invest on several issues including parameter selection (for parameter optimization), error characterization on both model and measurements, …
- For carbon flux upscaling with model parameter optimization, the combination of multiple data streams has proven to provide larger constraint than using one data stream only; however difficulties arise in combining different observation types, in particular to define the relative weight of each data stream in the objective function that you minimize

Several recommendations as well as future steps can be drawn:

- There is a need to better quantify the information content of recent and new data stream that could be used in a DA framework (fluorescence, soil carbon, soil surface humidity,…)
- We need to have large number of in-situ observations that cover different climate regimes, plant species and soil types to derive robust model parameters or model state variables.
- We need to be able to account for the history of the site when focusing on the carbon cycle, which require land cover change history and for forests thinning events.
- There is a need to account for carbon stocks, especially in the soil, when optimizing model parameters in order not bias the retrieval of decomposition rates.

6. References

Aubinet, M., Vesala, T., and Papale, D.: Eddy covariance: a practical guide to measurement and data analysis, Springer Science & Business Media, 2012.

Baldauf, M., Förstner, J., Klink, S., Reinhardt, T., Schraff, C., Seifert, A., Stephan, K., and Wetterdienst, D.: Kurze Beschreibung des Lokal-Modells Kürzestfrist COSMO-DE (LMK) und seiner Datenbanken auf dem Datenserver des DWD, Deutscher Wetterdienst, Geschäftsbereich Forschung und Entwicklung, Offenbach, Germany, 2009.

Baldocchi,, D. (2008), 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, Aust. J. Bot., 56, 1–26, doi:10.1071/BT07151.

Bogena, H. R., M. Herbst, J. A. Huisman, U. Rosenbaum, A. Weuthen, and H. Vereecken (2010), Potential of Wireless Sensor Networks for Measuring Soil Water Content Variability, Vadose Zone Journal, 9(4), 1002-1013, doi:10.2136/vzj2009.0173.

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., Lawrence, D. M., and Swenson, S. C.: Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, J Geophys Res-Biogeo, 116, 2011.

Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Global Change Biology 11, 335–355. doi:10.1111/j.1365-2486.2005.00897.x

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble Kalman filter, Monthly Weather Review, 126(6), 1719-1724.

Carvalhais, N.*, Reichstein, M.*, Seixas, J., Collatz, G.J., Pereira, S.J., Berbigier, P. Carrara, A., Granier, A., Montagnani, L., Papale, D., Rambal, S., Sanz, M. J. and Valentini, R. (2008) *Implications of the carbon cycle steady state assumptions for biogeochemical modeling performance and inverse parameter retrieval.* Global Biogeochemical Cycles, 22, GB2007, doi:10.1029/2007GB003033

F. Campolongo, J. Cariboni, A. Saltelli, An effective screening design for sensitivity analysis of large model, Environmental Modelling & Software, Volume 22, Issue 10, October 2007, Pages 1509–1518.

Evensen, G. (1994), Sequuential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forcast error statistics, Journal of Geophysical Research-Oceans, 99(C5), 10143-10162, doi:10.1029/94jc00572.

FAO, IIASA, ISRIC, ISSCAS, and JRC: Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, 2012.

Han, X., Li, X., Franssen, H. J. H., Vereecken, H., and Montzka, C.: Spatial horizontal correlation characteristics in the land data assimilation of soil moisture, Hydrology and Earth System Sciences, 16, 1349-1363, 2012.

Han, X. J., Franssen, H. J. H., Montzka, C., and Vereecken, H.: Soil moisture and soil properties estimation in the Community Land Model with synthetic brightness temperature observations, Water Resour Res, 50, 6081-6105, 2014.

Hollinger, D.Y., Richardson, A.D., 2005. Uncertainty in eddy covariance measurements and its application to physiological models. Tree Physiol. 25, 873-885.

Hou, Z. S., Huang, M. Y., Leung, L. R., Lin, G., and Ricciuto, D. M.: Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model, J Geophys Res-Atmos, 117, 2012.

Jung, M., Reichstein, M., Margolis, H. a., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., et al.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res. Biogeosciences, 116, 2011.

Kaminski, T., W. Knorr, M. Scholze, N. Gobron, B. Pinty, R. Giering, and P.-P. Mathieu (2012), Consistent assimilation of MERIS FAPAR and atmospheric CO_2 into a terrestrial vegetation model and interactive mission benefit analysis, *Biogeosciences*, 9(8):3173-3184, 2012.

Kuppel, S., P. Peylin, F. Chevallier, C. Bacour, F. Maignan, and A.D Richardson (2012), Constraining a global ecosystem model with multi-site eddy-covariance data, *Biogeosciences*. 9, 3757-3776.

Kuppel, S., P. Peylin, F. Maignan, F. Chevallier, G. Kiely, L. Montagnani, and A. Cescatti (2014), Model–data fusion across ecosystems: from multi-site optimizations to global simulations, *Geosci. Model Dev. Discuss.*, 7, 2961-3011, doi:10.5194/gmdd-7-2961-2014

Kustas, W. and Anderson, M.: Advances in thermal infrared remote sensing for land surface modeling, Agr Forest Meteorol, 149, 2071-2081, 2009.

Liu, S. M., Xu, Z. W., Wang, W. Z., Jia, Z. Z., Zhu, M. J., Bai, J., and Wang, J. M.: A comparison of eddy-covariance and large aperture scintillometer measurements with respect to the energy balance closure problem, Hydrology and Earth System Sciences, 15, 1291-1306, 2011.

Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser (2005a), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, Advances in Water Resources, 28(2), 135-147, doi:10.1016/j.advwatres.2004.09.002.

Moradkhani, H., K. L. Hsu, H. Gupta, and S. Sorooshian (2005b), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water Resources Research, 41(5), doi:10.1029/2004wr003604.

MORRIS (Jr) J.H., PRATT V.R., 1970, *A linear pattern-matching algorithm*, Technical Report 40, University of California, Berkeley.

Oleson, K., Lawrence, D. M., Bonan, G., Drewniak, B., Huang, M., Koven, C. D., Levis, S., Li, F., Riley, W. J., Subin, Z. M., Swenson, S. C., Thornton, P. E., Bozbiyik, A., Fisher, B. E. A., Kluzek, E., Lamarque, J. F., Lawrence, P. J., Leung, L. R., Lipscomb, W., Muszala, S., Ricciuto, D. M., Sacks, W., Sun, Y., Tang, J., and Yang, Z.-L.: Technical Description of

version 4.5 of the Community Land Model (CLM), Ncar Technical Note NCAR/TN-503+STR, National Center for Atmospheric Research, Boulder, CO, 422 pp., 2013. 2013.

Pauwels, V.R.N., A. Balenzano, G. Satalino, H. Skriver, N.E.C. Verhoest, and F. Mattia, Optimization of soil hydraulic model parameters using Synthetic Aperture Radar data: an integrated multidisciplinary approach, IEEE Transactions on Geoscience and Remote Sensing, 47(2), 455-467, 2009.

Raupach, M.R., Rayner, P.J., Barrett, D.J., DeFries, R.S., Heimann, M., Ojima, D.S., Quegan, S., Schmullius, C.C., 2005. Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. Global Change Biology 11, 378–397.

Richardson, A.D., Hollinger, D.A., Burba, G.G., Davis, K.J., Flanagan, L.B., Katul, G.G., Munger, J.W., Ricciuto, D.M., Stoy, P.C., Suyker, A.E., Verma, S.B., Wofsy, S.C., 2006. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. Agric. Forest Meteorol. 136, 1-18.

Saatchi S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and A. Morel, 2011. Benchmark map of forest carbon stocks in tropical regions across three continents. Proc. Natl. Acad. Sci. U. S. A., 108, 9899-904.

Santaren, D., P. Peylin, C. Bacour, P. Ciais, and B. Longdoz, (2014), Ecosystem model optimization using in-situ flux observations: benefit of Monte-Carlo vs. variational schemes and analyzes of the year-to-year model performances, *Biogeosciences Discuss.*, 10(100), 18009–18064.

Schwinger, J., Kollet, S. J., Hoppe, C. M., and Elbern, H.: Sensitivity of Latent Heat Fluxes to Initial Values and Parameters of a Land-Surface Model, Vadose Zone J, 9, 984-1001, 2010.

Tarantola, A. (2005), Inverse problem theory and Methods For Model Parameter Estimations, Society for Industrial and Applied Mathematic, Philadelphia, 342pp.

Thurner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D., Shvidenko, A., Kompter, E., Ahrens, B., Levick, S. R. and Schmullius, C.: Carbon stock and density of northern boreal and temperate forests, Glob. Ecol. Biogeogr., 23, 297–310, 2014.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, and G. Schoups (2013), Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, Advances in Water Resources, 51, 457-478, doi:10.1016/j.advwatres.2012.04.002.

Walker, A. P., Aranda, I., Beckerman, A. P., Bown, H., Cernusak, L. A., Dang, Q. L., Domingues, T. F., Gu, L., Guo, S., Han, Q., Kattge, J., Kubiske, M., Manter, D., Merilo, E. M., G. , Porte, A., Scales, J. C., Tissue, D., Turnbull, T., Warren, C., Wohlfahrt, G., Woodward, F. I., and S.D., W.: A Global Data Set of Leaf Photosynthetic Rates, Leaf N and P, and Specific Leaf Area. Data set, Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1224, 2014.

Weiss, M., F. Baret, S. Garrigues, and R. Lacaze (2007), LAI and fAPAR CYCLOPES global products derived from VEGETATION. Part 2: validation and comparison with MODIS collection 4 products, *Remote Sens. Environ.*, 110(3): 317-331, 10.1016/j.rse.2007.03.001.

Whitaker, J. S., and T. M. Hamill (2012), Evaluating Methods to Account for System Errors in Ensemble Data Assimilation, Monthly Weather Review, 140(9), 3078-3089, doi:10.1175/mwr-d-11-00276.1.