

Carbon cycle data assimilation with a generic phenology model

W. Knorr¹, T. Kaminski², M. Scholze¹, N. Gobron³, B. Pinty³, R. Giering², P.-P.

Mathieu⁴

¹QUEST – Dep. of Earth Sciences, Wills
Memorial Building, University of Bristol,
Bristol, BS8 1RJ, U.K.

²FastOpt GmbH, Schanzenstrae 36, 20357
Hamburg, Germany

³European Commission, DG Joint Research
Centre, Institute for Environment and
Sustainability, Global Environment Monitoring
Unit, TP 272, via E. Fermi, 21020 Ispra (VA),
Italy

⁴European Space Agency, Earth Observation
Science & Applications, Via Galileo Galilei,
Casella Postale 64, 00044 Frascati (Rm), Italy

Abstract. Photosynthesis by terrestrial plants is the main driver of the global carbon cycle, and the presence of actively photosynthesizing vegetation can now be observed from space. However, challenges remain when translating remotely sensed data into carbon fluxes. One reason is that the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), which documents the presence of photosynthetically active vegetation, relates more directly to leaf development and leaf phenology than to photosynthetic rates. Here, we present a new approach for linking FAPAR and vegetation-to-atmosphere carbon fluxes through variational data assimilation. The scheme extends the Carbon Cycle Data Assimilation System (CCDAS) by a newly developed, globally applicable and generic leaf phenology model, which includes both temperature and water-driven leaf development. CCDAS is run for seven sites, six of them included in the FLUXNET network. Optimization is carried out simultaneously for all sites against 20 months of daily FAPAR from the Medium Resolution Imaging Spectrometer (MERIS) on-board the European Space Agency's ENVISAT platform. 14 parameters related to phenology and 24 related to photosynthesis are optimized simultaneously and their posterior uncertainties computed. We find that with one parameter set for all sites, the model is able to reproduce the observed FAPAR spanning boreal, temperate, humid-tropical and semi-arid climates. Assimilation of FAPAR has led to reduced uncertainty ($>10\%$) of 10 of the 38 parameters, including 1 parameter related to photosynthesis, and a moderate reduction in NPP uncertainty. The approach can easily be extended to regional or global studies and to the assimilation of further remotely sensed data sources.

1. Introduction

The terrestrial biosphere does not only respond to climate fluctuations and change, it is an active part of the Earth's climate system. By taking up or releasing carbon, it directly impacts on the levels of CO₂ in the atmosphere, influencing climate [Friedlingstein *et al.*, 2006]. Therefore, understanding how CO₂ fluxes between the atmosphere and the terrestrial biosphere respond to interannual fluctuations of today's climate is crucial for sound projections of future climate change [Ciais *et al.*, 2005; Knorr *et al.*, 2005b; Zeng *et al.*, 2005]. Capturing changes of the terrestrial biosphere by satellites is an especially interesting method, because of its global coverage. The presence of healthy vegetation can be captured well from space, because it exhibits a strong contrast in reflectance between the visible and the near-infrared part of the solar spectrum [Verstraete *et al.*, 1996]. This feature is robust across all land plants because it has its roots in an adaptive mechanism: visible light is photosynthetically active, while near-infrared light is not, and therefore its absorption is avoided to reduce overheating and transpirational loss [Jones, 1983]. This robustness of spectral features has led to the design of various vegetation indices [Deering *et al.*, 1975; Verstraete and Pinty, 1996; Huete, 2007], but the quantity best related to the phenomenon just discussed is the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR). In the context of this study, it is defined as the amount of radiation in the photosynthetically active range (0.3-0.5 μ m) absorbed by healthy leaves or conifer needles divided by the total amount of radiation absorbed at the surface (i.e. one minus albedo in that spectral range) [Pinty *et al.*, 2009]. Thus, an optically deep canopy composed entirely of leaves would be able to reach a theoretical maximum close to 1, while no vegetation always amounts to a value of 0.

This study is based on the assumption that agreement between satellite-derived FAPAR and FAPAR computed by the vegetation model assures that given the same illumination conditions, the modelled photon absorption by leaves is consistent with what was estimated during the satellite retrieval. However, it must be borne in mind that the proper interpretation of FAPAR data requires understanding of the various physical processes that impact the reflected light on its course through the atmosphere and the canopy before it arrives at the sensor. This understanding has been embedded in radiative transfer models that are then used in inversion methods to retrieve vegetation parameters from remote sensing data. Various mathematical tools have been developed that minimize the effects due to the scattering by atmospheric particles, the brightness of soils and the changing geometry of illumination and observation.

Current operational remotely sensed FAPAR products are mainly derived from medium-resolution satellite instruments to provide regional and global operational FAPAR products at a variety of spatial and temporal scales. They have been designed for various sensors, such as the Moderate Resolution Imaging Spectroradiometer (MODIS, [Knyazikhin *et al.*, 1998a]), the Multiangle Imaging Spectro-Radiometer (MISR, [Knyazikhin *et al.*, 1998b]), the Sea-viewing Wide Field-of-view Sensor (SeaWiFS; Gobron *et al.* [2002]) and the Medium Resolution Imaging Spectrometer (MERIS; Gobron *et al.* [2008]), as part of various projects, such as LANDSAF [Roujean and Bréon, 1995] and CYCLOPES [Baret *et al.*, 2007], and as a derived product of broadband surface albedo [Pinty *et al.*, 2007].

FAPAR products have been used to monitor large-scale changes in vegetation status [Gobron *et al.*, 2005] and have then been related to large-scale changes in land-atmosphere CO₂ fluxes via algorithms for net primary productivity (NPP) [Cao *et al.*, 2004] or by validation with results from vegetation modelling [Zeng *et al.*, 2005; Knorr *et al.*, 2005a, 2007]. The latter approach

raised the prospect of eventually identifying all major processes at the ecosystem level that are responsible for the climate response and including them in an assimilation system, open up the way for improving future assessments of climate change impacts. This has been taken a step further by the Carbon Cycle Data Assimilation System (CCDAS; *Rayner et al.* [2005]). Here, observational evidence for interannual changes in atmosphere-biosphere carbon fluxes is assimilated into a terrestrial vegetation model through constraining prior uncertainty ranges of process parameters. This approach allows using data that are only rather indirectly related to the process studied, in the case cited concentrations of atmospheric CO₂ at remote flask sampling sites, which would normally exclude them from being used for anything but validation of model results.

This advantage of using indirect information for constraining mathematical process descriptions is also valid for the use of FAPAR for constraining rates of carbon fluxes and even photosynthesis. Even though FAPAR has been used in the past to infer NPP by multiplying with incident photosynthetically active radiation (PAR) and some efficiency factor [*Monsi and Saeki*, 1953; *Prince*, 1991; *Ruimy et al.*, 1994], the presence of leaves is determined not by the rate of photosynthesis, but by the adaptation of the plant to environmental stresses such as frost or drought [*Woodward*, 1987], a process described as leaf phenology. Therefore, CCDAS has used FAPAR not as an input quantity for an efficiency NPP model, but as a constraint on a small number of model parameters independently for each grid cell of the global model. This has been carried out in a first assimilation stage [*Knorr and Schulz*, 2001], which was not included in the overall estimation of parameter uncertainties [*Rayner et al.*, 2005] and their propagation to diagnostic and prognostic uncertainties of carbon fluxes [*Scholze et al.*, 2007].

The main aim of the present study is therefore to report on progress in the further development of CCDAS, where the assimilation of FAPAR data has been fully integrated into the variational stage of the system. What is meant here is that a large set of parameters is optimized for a multitude of locations simultaneously by minimizing a global cost function, and the second derivative of the cost function is used to infer improvements in parameter accuracy from the prior to the posterior estimate. Details of the approach are given in Section 3. The second aim is to assess in more general what information is added by including FAPAR into a modelling framework aimed at simulating NPP in addition to information from climate, land use and soils data. This second aim is motivated by the widely used efficiency approach for modelling carbon fluxes from FAPAR [*Field et al.*, 1998; *Zhao et al.*, 2005; *Turner et al.*, 2009] (i.e. where FAPAR is used as an input parameter and the models therefore can run only for time periods where such observations are available), and by the existence of satellite-derived net (NPP) and gross primary production (GPP) products [*Zhao et al.*, 2005; *Sims et al.*, 2008].

2. Notation

We use a notation very similar to that of *Scholze et al.* [2007].

x Vector of control variables

x_0 Prior values of control variables

C_0 Covariance of the associated uncertainty

d Observations

C_d Covariance of the associated uncertainty

x_{post} Optimal (posterior) values of control variables

C_{post} Covariance of the associated uncertainty

y Diagnostic/prognostic quantity of interest (target quantity), e.g. a net flux

C_y Covariance of the associated uncertainty

$M(x)$ Model operated as a mapping from control variables onto counterparts of observations.

$N(x)$ Model operated as a mapping from control variables onto target quantity.

N' Its linearization around x_{post} .

$J(x)$ Cost function

$H(x)$ Hessian of J , i.e. the matrix composed of its second partial derivatives.

3. Description of CCDAS

The set-up, data and models used in CCDAS have been described by *Scholze* [2003], *Rayner et al.* [2005], and *Scholze et al.* [2007] to which we refer for details regarding the formulation. In brief, BETHY, the core CCDAS model, is a process-based model of the terrestrial biosphere [*Knorr*, 2000]. It simulates carbon assimilation and plant and soil respiration embedded within a full energy and water balance and phenology scheme. BETHY is a fully prognostic model, and is thus able to predict the future evolution of the terrestrial carbon cycle under a prescribed climate scenario. Global vegetation is mapped onto 13 plant functional types (PFT) based on *Wilson and Henderson-Sellers* [1985] (see Table 1 for the PFTs covered by this study). Each grid cell of arbitrary size can contain up to three different PFTs, with the amount specified by their fractional coverage. As mentioned, all previous CCDAS studies have used a two-stage inversion procedure, where the first stage uses the full BETHY model to assimilate fields of FAPAR derived from satellite data, thereby optimizing parameters controlling soil moisture and phenology. The second stage uses a reduced version of BETHY with no phenology scheme and

no water balance, to assimilate atmospheric CO₂ concentration observations from a global station network [*GLOBALVIEW-CO₂*, 2004]. This simplified form of the model uses the leaf area index (LAI) and plant available soil moisture fields provided by the first stage after optimization.

The assimilation of FAPAR data in CCDAS [*Knorr and Schulz*, 2001] is based on minimization of the difference between satellite and model-derived FAPAR. Within BETHY, FAPAR is calculated as the vertical integral of absorption of photosynthetically active radiation by healthy green leaves divided by the difference between the incoming and outgoing radiation flux at the top of the canopy. This integration is carried out by a two-flux scheme, which takes into account soil reflectance, solar angle and amount of diffuse radiation. Equating satellite and model FAPAR means that given the same illumination conditions, the same number of photons enter the photosynthetic mechanism of the vegetation, even if some of the assumptions differ between BETHY and the model used to derive FAPAR [*Gobron et al.*, 2000]. It also means that FAPAR in the model is defined only with respect to the absorption by photosynthesizing plant parts [*Pinty et al.*, 2009], which is consistent with the definition used for deriving the MERIS FAPAR product.

The second stage then allows the rigorous propagation of uncertainties as described by *Kaminski et al.* [2002, 2003]; *Rayner et al.* [2005] and demonstrated by *Kaminski et al.* [2002]; *Rayner et al.* [2005]; *Scholze et al.* [2007]. It uses a probabilistic framework, described in detail by *Tarantola* [1987] or *Enting* [2002], who also gives an exhaustive overview on applications to biogeochemistry.

The state of information on a specific physical quantity is conveniently formulated in terms of a probability density function (PDF). The prior information is quantified by a PDF in the space of control variables (here: process parameters of BETHY and the initial atmospheric

CO₂ concentration), and the observational information by a PDF in the space of observations. Their respective means are denoted by x_0 and d and their respective covariance matrices by \mathbf{C}_0 and \mathbf{C}_d . Note that \mathbf{C}_d has to account for uncertainties in the observations and uncertainties from errors in simulating their counterpart. We approximate the posterior PDF by a Gaussian with mean x_{post} and covariance matrix \mathbf{C}_{post} . The mean is the minimum of the following cost function:

$$J(x) = \frac{1}{2}[(M(x) - d)^T \mathbf{C}_d^{-1}(M(x) - d) + (x - x_0)^T \mathbf{C}_0^{-1}(x - x_0)] , \quad (1)$$

where $M(x)$ denotes the model operated as a mapping of the control variables onto simulated counterparts of the observations. In practice, the minimization of J is performed iteratively by a gradient algorithm. The search direction is determined via the gradient of J , evaluated by adjoint code. The use of adjoint model code greatly enhances computational performance of the non-linear optimization.

We approximate the covariance matrix of the model parameters as

$$\mathbf{C}_{post}^{-1} = \mathbf{H}(x_{post}) , \quad (2)$$

where $\mathbf{H}(x_{post})$ denotes the Hessian matrix of J , i.e. the matrix composed of its second partial derivatives $\frac{\partial^2 J}{\partial x_i \partial x_j}$. Since the dimension of x_{post} never exceeds a few hundred, it is computationally feasible to evaluate the full Hessian by running efficient second derivative code.

The inverse step is followed by a second step, the estimation of a diagnostic or prognostic target quantity y . The corresponding PDF is approximated by a Gaussian with mean

$$y = N(x_{post}) \quad (3)$$

and covariance

$$\mathbf{C}_y = \mathbf{N}'(x_{post}) \mathbf{C}_{post} \mathbf{N}'(x_{post})^T + \mathbf{C}_{y,mod} , \quad (4)$$

where $N(x)$ is the model operated as a mapping of the control variables onto the target quantity. In other words, the model is expressed as a function of the vector of its parameters x and returns a vector of quantities of interest, for example the rate of photosynthesis at some desired time step. $\mathbf{N}'(x_{post})$, the Jacobian matrix of N , is its linearization around x_{post} , and $\mathbf{C}_{y,mod}$ is the uncertainty in the simulation of y resulting from errors in the model. In the extreme, hypothetical case that the model was perfect, only the first term would contribute to \mathbf{C}_y . On the other hand, if the control variables were known to perfect accuracy, only the second term would contribute to \mathbf{C}_y .

The minimization of Equ. (1) and the propagation of uncertainties are implemented in a normalized parameter space with Gaussian prior. The normalization is such that parameter values are specified in multiples of their standard deviation, i.e. \mathbf{C}_0 is the identity (for details see *Kaminski et al.* [1999]; *Rayner et al.* [2005]). We assume that Hessian Eigenvalues less than 1 reflect small-scale noise. To remove this noise Hessian Eigenvalues around 1 are set to 1 as in the procedure detailed by *Rayner et al.* [2005].

The basic set-up of the previously used and the newly developed CCDAS are shown in Fig. 1. The previous scheme consists of two stages, and satellite-derived FAPAR are only assimilated in Stage 1. Since uncertainties of parameters and diagnostic quantities (such as NPP) are only evaluated in Stage 2, no information on the reduced uncertainty after assimilating FAPAR data is available. The purpose of this two-stage process is to reduce computational demand for the optimization in Stage 2, which is run only with part of the full BETHY model of Stage 1. Instead, the state variables of the phenology and hydrology model are passed from Stage 1 to Stage 2, where the reduced “carbon-BETHY” lacks those two components. The quantity assimilated in Stage 2 is atmospheric CO_2 , which requires an atmospheric transport model (TM2) as

well as additional background CO₂ fluxes. Those background fluxes, which enter directly into TM2, are not required for the new set-up, but would be if this were applied at the global scale with assimilation of CO₂ concentrations. This is currently under development.

FAPAR in the new set-up is directly assimilated into “CCDAS BETHY”, which includes the hydrology model of the previous full BETHY (with some slight modification, see below), and the newly developed phenology model which replaces the one of full BETHY [Knorr, 2000]. Another modification, which is not depicted in Fig. 1, is that BETHY can now be switched between modes where it runs for a multitude of sites simultaneously (as in this study, see Section 5.3)), globally, or simultaneously at sites and globally.

By including the FAPAR assimilation into the previous Stage 2, the derivative-based framework can now be used to propagate uncertainties in observed FAPAR back to uncertainties in control variables and then forward to uncertainties in target quantities. This change, however, requires some modification to the model as it was present in full BETHY, which is the reason why a new phenology scheme had to be developed. The requirement is that the simulated FAPAR (and, with it, the cost function in Equ. (1)) and the carbon fluxes (see Equ. (3)) are at least twice differentiable functions of all process parameters. As it is explained below, this was not the case for the phenology scheme of the previous full BETHY. All derivative code of Stage 2 or the new CCDAS is generated from the model code by the automatic differentiation tool Transformation of Algorithms in Fortran (TAF) [Giering and Kaminski, 1998].

The following changes to the BETHY version of Knorr [2000] were necessary: First, the phenology model was replaced by a new scheme described in Section 4. Second, the soil evaporation model was simplified such that soil evaporation, E_s , happens at the following rate:

$$E_s = E_{s,pot} * w_s / w_{s,max}, \quad (5)$$

where $E_{s,pot}$ is the potential rate determined by the net radiation at the soil surface, w_s the amount of water in a second soil water bucket, and $w_{s,max}$ its maximum value. The modified BETHY works with two overlapping buckets, both of which have exactly the same inputs and outputs, but the shallower bucket, which overflows earlier, determines soil evaporation via Equ. (5), whereas the original, deeper bucket (maximum value $W_{s,max}$) determines transpiration rates. We set $w_{s,max}$ to the lower of 5 mm and $W_{s,max}$ (see Section 5.2). Third, the size of the large bucket is redistributed between trees and herbaceous/shrub PFTs such that the average $W_{s,max}$ of the herbaceous PFTs at each grid cell is 30% of the average of the tree PFTs.

4. A generic global phenology scheme

The purpose of this section is to explain the development of a new generic phenology scheme that captures all major phenology types of the global terrestrial biosphere with one set of equations. Its parameters are described in Table 2. Differences between functional types are formulated entirely through differences in parameters, as shown in Table 3. Other required features of the scheme are explained next.

The remainder of this section is laid out as follows: Section 4.1 explains the consequences of using the derivative approach for the design of the phenology scheme. Section 4.2 explains the approach to sub-grid variability necessary to ensure differentiability, and introduces parameters for temperature and light dependent growth triggers. These are discussed in detail in Section 4.4, following Section 4.3, which introduces the differential equation describing the time evolution of the LAI. Finally, Section 4.5 discusses water and other limitations and their implementation via a maximum LAI, with a single parameter allowing for different water-use strategies between PFTs.

4.1. Requirements

Currently available phenology schemes for global vegetation models have some issues that make them unsuitable for use in CCDAS. Firstly, the present phenology scheme of BETHY relies heavily on NPP to determine LAI under drought conditions [Knorr, 2000]. Several tests with CCDAS have shown that this leads to oscillations around an optimal point where LAI and NPP are in equilibrium. Those oscillations then cause sudden jumps in the cost function with extremely small changes in the control parameters. An alternative approach using soil water content directly to drive drought-limited LAI was not considered, because the amount of soil water sufficient to support a certain LAI depends on the local evaporative demand. More complex schemes such as the phenology scheme of SILVAN [Kaduk and Heimann, 1996] or logro-P (C. Reick, *pers. comm.*, implemented in Raddatz *et al.* [2007]) use a number of switches, or “triggers”, to change the state of the vegetation between dormant, growth or senescent. With such a formulation involving discrete state variables, however, the cost function in Equ. (1) cannot be differentiated where one of those states changes. One goal of the present work is therefore to find a globally applicable but reasonably simple phenology scheme that should be differentiable everywhere with as few as possible sudden changes.

In the following we present such a newly devised phenology scheme with state variables that depend on the process parameters in differentiable form. The complexity of the formulation and the number of process parameters have been kept low. Nevertheless, the scheme allows for both temperature and water limitation and is able to represent the major global phenology types, namely cold-deciduous, warm-deciduous, cold-evergreen, moist-evergreen, grass and annual crop. This constitutes an important advance over similar global schemes that have been derived

from remotely sensed information, but either consider only temperature limited phenology, or use only soil moisture without considering evaporative demand [Botta *et al.*, 2000].

4.2. A general spatial approach

The phenology models cited above use triggers that set vegetation instantly from a dormant to an active, and again to a senescent stage. However, in reality over an area the size of a model grid cell, such transitions will happen at different times because either the environmental conditions or the exact point of the trigger vary. Apart from being unrealistic, the sudden change of state will create a non-differentiable dependency of the state variables on the control parameters (see Section 4.4).

Here we assume that spatial variability within a grid cell is entirely the result of differences in the threshold parameter defining the trigger, effectively subsuming impacts of small-scale climatic variability under the same. This parameter is assumed to have a Gaussian probability distribution in space. There are two of those threshold parameters: \tilde{T}_ϕ and \tilde{t}_c . It is important to note that the transition to the active state requires both $\tilde{T}_\phi > T$ and $\tilde{t}_c > t_d$, where T is a temperature and t_d length of day. The tilde () denotes that these are parameters or state variables of individual plants within a grid cell. As shown next, these parameters are integrated over their probability distribution, replacing the integration across the space of the grid cell.

Before proceeding to the spatial integration, we define a generic differential equation in time for the LAI of individual plants, $\tilde{\Lambda}(t)$:

$$\frac{d\tilde{\Lambda}(t)}{dt} = \begin{cases} f_1, & \text{if } T \geq \tilde{T}_\phi \text{ and } t_d \geq \tilde{t}_c, \\ f_2, & \text{else,} \end{cases} \quad (6)$$

where f_1 and f_2 are some arbitrary functions of the state of the vegetation.

In this discrete formulation, the response of LAI to changes in \tilde{T}_ϕ or \tilde{t}_c is usually non-differentiable at the threshold. The continuous version of Equ. (6), which is valid for the spatially integrated LAI, $\Lambda(t)$, resolves to an integral over the Gaussian probability density functions (PDF), p and q , of the two trigger variables.

$$\frac{d\Lambda(t)}{dt} = \int_{-\infty}^T \int_{-\infty}^{t_d} f_1 p(\tilde{T}_\phi) q(\tilde{t}_c) d\tilde{T}_\phi d\tilde{t}_c + \int_T^{+\infty} \int_{t_d}^{+\infty} f_2 p(\tilde{T}_\phi) q(\tilde{t}_c) d\tilde{T}_\phi d\tilde{t}_c, \quad (7)$$

The spatial PDF p is characterized by a mean T_ϕ and its standard deviation T_r , while the mean of q is t_c and the standard deviation t_r . All four are CCDAS control parameters. Note the distinction between these two spatial PDFs and the fact that their four parameters have again PDFs in the Bayesian sense in the same way as all other control parameters.

Since f_1 and f_2 depend themselves neither on \tilde{T}_ϕ nor on \tilde{t}_c , the previous expression simplifies to

$$\frac{d\Lambda(t)}{dt} = f f_1 + (1 - f) f_2 \quad (8)$$

with

$$f = \int_{-\infty}^T p(\tilde{T}_\phi) d\tilde{T}_\phi = \Phi\left(\frac{T - T_\phi}{T_r}\right), \quad (9)$$

where Φ is the cumulative normal distribution. f is the fraction of plants within the proportion of a grid cell occupied by each PFT that are actively growing or maintaining leaves.

4.3. Time evolution of LAI

Describing the generic formulation for the time evolution of LAI of a single plant – which is then integrated spatially via Equ. (8) – requires definition of f_1 (for plants in their growing season) and f_2 (for senescent plants). For f_1 , we assume the simplest formulation that satisfies the following two conditions: leaf growth starts immediately and is not limited by substrate availability, such as LAI itself; and growth stops if a target LAI is reached that is in balance with

the environmental limitations, described as Λ_{max} . These conditions are met by the following formulation:

$$f_1 = \xi(\Lambda_{max} - \Lambda), \quad (10)$$

where ξ is a linear growth constant describing the increase in LAI per time unit shortly after bud burst. This rate is chosen to be independent of carbon gains (NPP), because initial leaf development relies on buds and reserves from the previous year [Kaduk and Heimann, 1996]. This formulation differs from those used in similar applications, such as logro-P or the one by Liu *et al.* [2008], where the initial growth is exponential resulting in a logistic function for the time integral under constant conditions. Equ. (10) results in a time dependence described by $\Lambda(t)/\Lambda_{max} = 1 - \exp(-\xi t)$ for $\Lambda(0) = 0$, which is linear for small t . The advantage of this approach is that it does not require setting a minimum LAI to set off growth, which was 0.1 in the work by Liu *et al.* [2008]. This would not work here, because Λ_{max} might be less than such a minimum value. The difference in approach can be explained by the fact that the work just cited is restricted only to temperature controlled phenology and does not include situations where either the temperature or the water balance only allows small values of LAI.

For those plants that are outside their growth stage, we again chose the simplest formulation that allows accommodating both deciduous and evergreen phenology:

$$f_2 = \Lambda/\tau_L \quad (11)$$

The new parameter τ_L , which is related to leaf longevity, describes how quickly leaves are shed, or whether they stay inactive until the next growing season. Deciduous vegetation will normally shed leaves (which includes leaves turning brown, see above) within days to weeks. Evergreen vegetation on the other side should have values at the order of a year or more.

We now consider evolution of the spatially integrated grid-cell average LAI. Inserting Equ. (10) and Equ. (11) into Equ. (8) yields:

$$\frac{d\Lambda(t)}{dt} = \xi[\Lambda_{max} - \Lambda(t)]f - \frac{\Lambda(t)}{\tau_L}(1 - f). \quad (12)$$

In order to find a convenient form for integrating this expression, we define

$$r = \xi f + (1 - f)/\tau_L \quad (13)$$

and

$$\Lambda_{lim} = \xi\Lambda_{max}f/r, \quad (14)$$

so that Equ. (12) takes the form:

$$\frac{d\Lambda(t)}{dt} = \xi\Lambda_{max}f - r\Lambda(t) \quad (15)$$

$$= r[\Lambda_{lim} - \Lambda(t)]. \quad (16)$$

As long as f and Λ_{max} (and therefore r and Λ_{min}) do not depend on t , the equation above has the following solution:

$$\Lambda(t + \Delta t) = \Lambda_{lim} - [\Lambda_{lim} - \Lambda(t)]e^{-r\Delta t} \quad (17)$$

Here it is sufficient to state that Λ_{max} depends on quantities that are updated either daily or every fews days, while f depends on daily values of temperature and day length. Therefore, the last equation can be used to integrate over a single daily time step of the phenology scheme, Δt . This mixture of analytical and numerical integration is not only highly efficient, it also ensures stability, i.e. it avoids negative Λ as long as Λ_{lim} and $\Lambda(0)$ are non-negative.

4.4. Temperature and day length requirements

Determining the date of leaf onset in cold-seasonal climates is often approached by the concept of growing-degree days (GDDs), defined as the sum of the daily mean temperatures minus a

threshold temperature, T_{GDD} , as long as this contribution is positive (see the discussion by *Botta et al.* [2000]). In the simplest approach, the parameters of the scheme would be the critical GDD (GDD_c), and T_{GDD} . GDD approaches attach the same weight to temperatures in the past, and therefore must be reset at certain times per year, for example at the beginning of January for sites in the northern temperate zone. For a global model, this might lead to complications when determining the appropriate date of the reset. Further, the LAI at a given day with a given GDD will be some positive number if $\text{GDD} \geq \text{GDD}_c$, but 0 if $\text{GDD} < \text{GDD}_c$. If GDD is close to GDD_c , then a small change in GDD_c can lead to a jump in the LAI at leaf onset. The LAI is therefore not always differentiable with respect to the parameter GDD_c . The same applies to T_{GDD} , as it changed GDD. Thus, the GDD concept produces non-differentiable dependencies of the state variables on the parameters.

Instead of the GDD sum, our approach uses a phenology determining temperature, T , defined as the normalized integral over the 2m air temperature, T_{2m} , with exponentially declining weights attached when going back into the past:

$$\begin{aligned} T(t) &= \frac{\int_{-\infty}^0 T_{2m}(t + \tilde{t}) e^{\tilde{t}/\tau_m} d\tilde{t}}{\int_{-\infty}^0 e^{\tilde{t}/\tau_m} d\tilde{t}} = \frac{1}{\tau_m} \int_{-\infty}^0 T_{2m}(t + \tilde{t}) e^{\tilde{t}/\tau_m} d\tilde{t} \\ &= \frac{1}{\tau_m} e^{-t/\tau_m} \int_{-\infty}^t T_{2m}(t') e^{t'/\tau_m} dt'. \end{aligned} \quad (18)$$

This is equivalent to an exponentially declining memory of the plants for the ambient temperature.

The remaining parameter is τ_m , the averaging period for T . Decreasing τ_m means that the temperature trigger uses more recent temperature data, so that the threshold is reached earlier and the vegetation period is shifted forward in the season. However, if τ_m becomes too short, T oscillates because short-term fluctuations in T_{2m} are not dampened any more. Because this

leads to instabilities in the optimization, the parameter is held fixed at a value of 30 days. This is not an undue limitation of the model, because the optimization already has the possibility to both extend the length of the growing season and shift it in time through changing both T_ϕ and t_c .

It is computationally favorable to bring Equ. (18) to an incremental form:

$$\begin{aligned}
 T(t + \Delta t) &= \frac{1}{\tau_m} e^{-(t+\Delta t)/\tau_m} \int_{-\infty}^{t+\Delta t} T_{2m}(t') e^{t'/\tau_m} dt' \\
 &= \frac{1}{\tau_m} e^{-\Delta t/\tau_m} e^{-t/\tau_m} \left(\int_{-\infty}^t T_{2m}(t') e^{t'/\tau_m} dt' + \int_t^{t+\Delta t} T_{2m}(t') e^{t'/\tau_m} dt' \right) \\
 &= e^{-\Delta t/\tau_m} T(t) + \frac{1}{\tau_m} e^{-(t+\Delta t)/\tau_m} \int_t^{t+\Delta t} T_{2m}(t') e^{t'/\tau_m} dt' \quad (19)
 \end{aligned}$$

If Δt is the time step of the model and T_{2m} constant over that period, then the expression simplifies to

$$T(t + \Delta t) = e^{-\Delta t/\tau_m} T(t) + T_{2m}(t)(1 - e^{-\Delta t/\tau_m}), \quad (20)$$

which allows continuous updating of T with only instantaneous values of T_{2m} . In practice, this scheme is implemented by setting $T = T_{2m}$ at the beginning of the model run and performing a sufficiently long spin-up.

The temperature condition for leaf growth is simply $T > \tilde{T}_\phi$. It is supplemented by an additional trigger for day length, t_d , with the condition $t_d > \tilde{t}_c$. Because the warmest period across the year is considerably later than the period of maximum daylight hours, this double condition has the effect that leaf onset is triggered by \tilde{T}_ϕ , and leaf shedding by \tilde{t}_c . As a reminder: the control parameters are T_ϕ and t_c (means), and T_r and t_r (standard deviations on a grid-cell).

According to *White et al.* [1997], of the PFTs considered here only cold-deciduous and cold-evergreen trees and shrubs have a day length trigger for leaf shedding (see Table 3). Grass PFTs do not have such a day length requirement, so that for those we set $t_c = 0$. Warm-evergreen and

warm-deciduous PFTs also do not have an explicit temperature requirement for growth because they are not actively protected against cold conditions. For these, we set $T_\phi = -\infty$. For all other PFTs, these are control parameters. The temperature trigger for the cold-deciduous and cold-evergreen (Table 1) is estimated to lie in the region of 10°C, except for the understory shrubs (PFT 8) where it would be somewhat lower, and only little above freezing for grasses (PFT 9,10).

4.5. Water and structural limitations

On a global scale, the main limiting factor on terrestrial plant growth is not temperature but water [Woodward, 1987]. Whenever photosynthesizing, plants lose water by transpiration through pore openings (“stomata”) in their leaves. This limitation together with any other limitation on leaf growth, here considered “structural”, is described by a single state variable, Λ_{max} .

If soil water is limiting, an increasingly negative soil water potential leads to a falling leaf potential in a complicated process involving root water uptake, xylem resistance to flow, and transpiration through the leaf stomata. If stomata close, leaves can retain water, but only to a degree that depends on its cuticular resistance, which in itself is dependent on the plant functional type (PFT). However, it is not possible to represent those complex mechanisms in a model designed for global-scale applications.

The scheme chosen here goes back to Woodward [1987], who used annual potential evapotranspiration and precipitation to derive water limited LAI on a global scale. To accommodate the shorter time scale of our model, we have modified his scheme by using daily actual transpiration and soil moisture instead of annual potential evapotranspiration and precipitation. However,

to have LAI react not to rapidly changing daily conditions but to the longer-term climatic state, the water limited LAI is averaged back in time using the same approach as for T (Equ. (18)).

Generally speaking, leaf development will stop and leaves will be shed if there is insufficient soil water for transpiration. At which level this happens exactly will be a function of various drought adaptations of the PFT concerned. Independent of the details, however, adaptation will determine how long the plant – at a given LAI – can survive with a given amount of soil moisture without rain. This time scale, τ_W , can serve as a universal parameter of water limitation. This defines a water-limited LAI, Λ_W through

$$E(\Lambda_W)\tau_W = W \quad (21)$$

where W is plant-available soil moisture. What is needed then is the total water loss after time τ_W as a function of leaf area.

To compute this water loss, we linearize the potential rate of transpiration, E , as a function of the LAI, Λ :

$$E(\Lambda) \approx \frac{\tilde{E}}{\tilde{\Lambda}} \Lambda. \quad (22)$$

\tilde{E} is the daily mean potential rate of transpiration last computed by the model at a LAI of $\tilde{\Lambda}$. This approximation is most accurate at low values of Λ and $\tilde{\Lambda}$, where net radiation of the leaf canopy, which drives evapotranspiration [*Jarvis and McNaughton*, 1986], can be assumed to scale linearly with LAI.

Combining Equ. (21) and Equ. (22) yields

$$\Lambda_W = \frac{W\tilde{\Lambda}}{\tilde{E}\tau_W}. \quad (23)$$

The parameter τ_W represents the expected length of drought periods. For $\tau_W \rightarrow 0$ the plant “expects” its water reserves to always be sufficient for continuing survival. In this case, $\Lambda_W \rightarrow$

∞ , meaning the plant has no explicit drought adaptation in its phenology. This is assumed for the cold-deciduous and cold-evergreen PFTs. For warm-evergreen plants, we expect the value for τ_W in the region of 1 year, and for grasses and warm-deciduous plants between one and two months. Water limitation is implemented separately for each PFT to reflect differences in the water use strategy, defined mainly by τ_W .

Observe also that for $\tilde{E} \rightarrow 0$ we have $\Lambda_W \rightarrow \infty$, since without evaporative demand the leaf area is not water limited, as it is the case for $\tau_W \rightarrow 0$. Since the LAI cannot grow indefinitely, it must be limited by other factors, such as light availability, nutrients and structure. These additional limitations are summarized into a single universal parameter $\hat{\Lambda}$ (cf. [Knorr, 2000]) and incorporated into the model via:

$$\tilde{\Lambda}_{max} = \nu(\hat{\Lambda}, \Lambda_W) \quad (24)$$

$\nu(x, y)$ is a smoothed minimum function defined by

$$\nu(x, y) = \frac{x + y - \sqrt{(x + y)^2 - 4\eta xy}}{2\eta} \quad (25)$$

with $\eta = 0.99$. $\tilde{\Lambda}_{max}$ is recomputed daily with daily values of the soil moisture, W , whereas \tilde{E} and $\hat{\Lambda}$ may be re-computed at longer intervals. This allows avoiding re-computing the diurnal cycle of photosynthesis and energy balance for every simulated day, while keeping a daily time step of phenology and water balance, to save computing time with the full BETHY model [Knorr, 2000].

Instead of $\hat{\Lambda}_{max}$, Equ. (10) uses Λ_{max} , its weighted time integration computed in the same way as T from T_{2m} (Equ. (9)), with the analogous definition:

$$\Lambda_{max}(t) = \frac{1}{\tau_s} e^{-t/\tau_s} \int_{-\infty}^t \tilde{\Lambda}_{max}(t') e^{t'/\tau_s} dt'. \quad (26)$$

Updating happens in the same way as described by Equ. (20) for T :

$$\Lambda_{max}(t + \Delta t) = e^{-\Delta t/\tau_s} \Lambda_{max}(t) + \hat{\Lambda}_{max}(t)(1 - e^{-\Delta t/\tau_s}). \quad (27)$$

The advantage of this scheme is that it has only one free parameter, τ_W , in addition to τ_s . Since changing τ_s might lead to instabilities of the optimization in a way similar to τ_m , this parameter is also held constant at value of 30 days.

5. Setup of the data assimilation system

This section explains the sites, the input data, the satellite FAPAR product used with the assimilation, and the parameterization of the model. A complete list of the parameters included in the assimilation is shown in Table 4. Some describe the site characteristics and are fixed, others are control parameters of CCDAS and are optimized during assimilation. CCDAS has the possibility of freely assigning control parameters either to one PFT subgrid cell at a particular station, or to all subgrid cells that represent the same PFT, or even to groups of PFTs. The parameter setup for the new phenology parameters uses this PFT grouping. Parameters of the previous CCDAS are also optimized and grouped in the same way as in previous work ([Rayner *et al.*, 2005; Scholze *et al.*, 2007]).

5.1. Site descriptions

Out of the total of 13 PFTs of the global version of CCDAS, seven distributed over eight sites occur in the assimilation study presented here (Table 1). On a global scale, the PFTs not considered (deciduous conifers, evergreen and tundra shrubs and wetlands) represent more marginal vegetation than the ones included.

Each site consists of a rectangular study area over one to several satellite pixels as described in Table 5. The last site shown in the table has been included for validation purposes and is

therefore excluded from the data assimilation exercise. The areas were chosen in such a way that they constitute homogeneous land cover as identified through Google Earth images. BETHY represents the vegetation of each site by two to three PFTs and a corresponding surface cover fraction shown in Table 6, where the remainder corresponds to bare ground. The water holding capacity of the soil ($W_{s,max}$) and the soil brightness class were extracted from the global version of BETHY. For the Hainich grass site, we use half the water holding capacity of the forest site, while for Manaus, we assume that deep roots cause the maximum available water storage to greatly exceed the values at the other sites [Nepstad *et al.*, 1994; Kleidon and Heimann, 1999]. The soil reflectance values for the brightness classes are 0.07 (dark) and 0.1 (medium) if the soil is wet, and 0.15 (dark) and 0.2 (medium) if the soil is dry.

The model is run once per PFT at each site, and the results are weighted according to the PFT fractions shown in Table 6.

5.2. Parameterization

The way phenology types are differentiated by choice of prior or fixed parameter values is summarized in Table 3. Fixed parameters always indicate that the corresponding limitation is deactivated, so the parameter does not need to be included in the optimization. Whether a plant is evergreen or deciduous is controlled either by τ_L or by τ_W .

In one case, a phenology type has a different parameterization depending on whether the vegetation is overstorey (PFT 4) or understorey (PFT 8). This reflects the different strategy of understorey plants, which need to develop leaves earlier to evade being shaded out. In general, however, parameters of different PFTs were grouped together whenever reasonable. (T_ϕ for PFTs 4 and 5, t_c of PFTs 4, 5 and 8, τ_L for PFTs 4, 8, 9, 10, and τ_W for PFTs 2, 9 10). The two grass PFTs (9, 10) always use the same control parameters for phenology as they are

only differentiated by photosynthetic pathway. We also know much less about the temperature control of grass phenology than we know about woody plants, where leaf on-set and shedding tend to happen within a well-defined period. Therefore, we assign a much larger uncertainty to T_ϕ and T_r for the two grass PFTs.

The complete list of parameters and how they are differentiated by PFT is shown in Table 4 together with their prior means and uncertainties. This table also represents the technical implementation of the control parameters, which differs from the model description in two cases: one is that τ_L is replaced by $k_L = 1/\tau_L$, which avoids division by small numbers in Equ. (12) and Equ. (13); and for τ_W , the control parameter is the natural logarithm, which in turn avoids negative numbers leading to negative LAI via Equ. (23). As the prior PDF is always Gaussian in the space of the control parameters, this transforms the prior PDF of both τ_L and τ_W accordingly. The logarithmic transformation has been described by *Rayner et al.* [2005].

For the two parameters representing within-pixel variability of growth triggers, T_r , and t_r , we only differentiate T_r by PFT, with one control parameter for trees and shrubs and one for grass. Here, we assume a larger spatial variability for the larger group containing trees and shrubs. There is only a single control parameter for t_r . Finally, $\hat{\Lambda}$, the structural LAI limit, is 5 as in *Knorr* [2000], and ξ is 0.5 per day, which describes rapid leaf sprouting where 97% of the maximum LAI has been reached after one week.

We assign rather large uncertainties in the region of 30 to 50% to most of the new parameters, considering the values of those parameters are known only to approximately. The remaining Parameters 1 to 24 were adopted from the previous version of CCDAS with prior values and uncertainties taken from *Scholze et al.* [2007]. The most important ones in the context of this

study are those controlling V_{max} (1-10), the maximum capacity of the enzyme that fixes the atmospheric CO_2 and makes it available for further metabolism.

5.3. MERIS FAPAR and input data

We assimilate daily data from the Level 2 FAPAR land product derived from the Medium Resolution Imaging Spectrometer (MERIS) of the European Space Agency (ESA) at the operational resolution of 1.2 km for the period June 2002 to September 2003. Square 15 by 15 pixel scenes have been processed that are centered at the position of the six sites previously introduced. We have used Google Earth imagery to identify areas of uniform cover type around the centers of the scenes and have selected rectangular sub-scenes of those original scenes, that were then spatially averaged. The extent of those sub-scenes (in km) and the number of valid daily data points after averaging are listed in Table 5. The fraction of days with valid data is highest in the tropics (32% for Maun) and lowest at the northern boreal site of Sodankylä (17%). We use an uncorrelated data uncertainty of 0.1 irrespective of how many pixels were used in the spacial averaging of the FAPAR pixels [Gobron *et al.*, 2008]. Thus, C_d in Equ. (1) contains only diagonal elements with values of either 0.1^2 , or ∞ if no data are available for the day and site concerned (in practice set to a very large value).

As described elsewhere [Knorr, 2000], the model is run with daily precipitation, minimum and maximum temperatures and incoming solar radiation. The data were generated through a combination of available monthly gridded and daily station data [R. Schnur, *pers. comm.*] by a method by Nijssen *et al.* [2001], using gridded data from the Summary of the Day Observations (Global CEAS), National Climatic Data Center and the latest updates of gridded data by Jones *et al.* [2001] and Chen *et al.* [2002] and using the available data nearest to the site. Input data related to soil and vegetation characteristics have already been described in Section 5.2.

6. Results

6.1. Calibration and Uncertainties

In the set-up described in Section 5.3, the minimization of Equ. (1) is carried out three times from three different starting points, and the procedure is repeated on two different computers. All six minimization procedures converge to the same minimum. The minimization starting from the prior value takes 30 iterations to reduce the cost function J from 822.4 to 468.3 and the norm of its gradient by more than seven orders of magnitude from 209.8 to 1.76×10^{-5} . The FAPAR observations lead to a substantial reduction of uncertainty (by 10% or more) in ten directions in parameter space.

Not surprisingly, most of the uncertainty reduction goes to parameters of the leaf phenology, namely: inverse leaf longevity (k_L , Parameters 34 and 35), the parameter describing adaptation to the expected length of dry periods (τ_W , Parameters 36 to 38), the temperature threshold for leaf onset (T_ϕ , Parameter 26 and 28), the spatial variability for grasses (T_r , Parameter 30) and the critical photoperiod (t_c , Parameter 31).

Except for grasses, parameters describing the spatial distribution of leaf onset or shedding are not or negligibly constrained (Parameters 29, 32), and so is the leaf growth parameter, ξ . We do not gain information on the phenology of understorey shrubs either (Parameter 27). The maximum leaf area index (Parameter 25) is also relatively little constrained. At high LAI, changes in LAI of the order of one or two can have little impact on FAPAR because FAPAR has already reached a value approaching to its theoretical maximum (close to one). As a result, the satellite-derived FAPAR is too large to resolve LAI values in the range of 4 to 5 [Gobron *et al.*, 1997].

Finally, a considerable reduction in uncertainty (by $> 10\%$) is only found for one physiology parameter, namely V_{max}^{25} for PFT 2, which only occurs at Maun. We note that Maun also has the largest number of data points (see Table 5).

6.2. Fit to observations

After assimilation, we found overall good agreement between data and observations considering the assumed data uncertainty (Figs. 2, 3). There are, however, a number of issues that remain to be addressed: the model overestimates the growing season length of the two grassland sites, the agreement for Manaus is less satisfactory, and there are frequent outliers in the data, notably around days 200 and 350 for Loobos. We suspect that the poor match at Manaus and the outliers are both due to remaining contamination by clouds not captured by the MERIS FAPAR algorithm, in particular cloud shadows, which are more difficult to detect than clouds themselves. We also note that the satellite data indicate an early spring greening at the Hainich forest site that is not captured by the model, even though the deviation is only slightly beyond the uncertainty range of the data. For the grass sites, it is interesting to note that the fit is better at Hainich (not included in the assimilation, root-mean-squared deviation of 0.165 at optimum) than at Aardhuis (deviation 0.267 at optimum), which was included. At Hainich, the model captures the summer drought in 2003 very well, better indeed than for the forest site, where the decline in FAPAR is less pronounced in both FAPAR and model. At both grass sites the algorithm overestimates the length of the growing season.

The data assimilation itself has led to an improvement of the fit at all sites, including the one where FAPAR data were not assimilated. The smallest improvement is found (in this order) for Loobos, Sodankylä and Zotino. In these cases, simulated FAPAR changes only slightly from prior to posterior case and prior agreement with the data is already good.

Table 4 shows both the prior and posterior parameter values, and the change in each parameter after optimization relative to the prior uncertainty. This relative change gives a measure of the extent to which the optimization has accepted our prior parameter estimates. A range of this value between ≈ -2 and $\approx +2$ corresponds to a change within the $\approx 95\%$ confidence interval and indicates agreement between model and prior estimate, while a value clearly outside this range indicates disagreement.

We observe that three parameters change by considerably more than two standard deviations of the prior PDF. These are Parameters 25 (maximum LAI, $\hat{\Lambda}$), 31 (critical photoperiod, t_c) and 36 (the water-use strategy parameter, τ_W , for tropical evergreen trees). For Parameters 25 and 31, we might have been too optimistic with our rather tight prior error margins.

For Parameter 36, however, the result suggests an extremely conservative strategy for evergreen tropical vegetation, apparently anticipating extremely long drought periods of almost three years. Examination of Fig. 3 (Manaus site) shows that even though FAPAR reaches very high values of > 0.8 , a large number of data points are below 0.5. This suggests that contamination by clouds and cloud shadows might be a serious problem at this site with its year-round humid climate. Obvious outliers to low FAPAR values were earlier identified for the Loobos site. Increasing τ_W during optimization had the effect of lowering Λ_W and thus improving the fit to the data. The optimal value for Λ_W at Manaus must therefore be considered unrealistic. For the warm-deciduous trees at Maun, however, the expected drought period is just over three months ($\tau_W = 112$ days), consistent with the length of the dry season at the site (150 days).

The problem with cloud contamination of FAPAR at Manaus also explains why Parameters 1 (V_{max}^{25} for PFT 1) and 25 ($\hat{\Lambda}$) were significantly downsized. The problem could have been

reflected in a larger error bar for the satellite FAPAR, which would have resulted in a larger weight given to the priors and thus in less change in the parameters.

For all sites we find that the vegetation period in cold climates is too long in the prior case. As a consequence, the optimization reduces T_ϕ for all PFTs, except for understory shrubs (PFT 8), and t_c is increased to 13.4 h by the optimization. *Oleksyn et al.* [1992] show even larger length-of-day thresholds in their study (14 hours and more), but referring to cessation of growth and not to the start of decline in LAI as it is the case with t_c . The latter implies that t_c needs to be less than the values they cite. Taken together, t_c seems consistent with independent ecophysiological data.

There is also a large reduction in Parameter 35 (k_L for PFT 5), which affects the three northernmost sites. As the deciduous vegetation dominates the seasonal amplitude, we expect that this parameter is more affected by the FAPAR observations outside the growing season. Such observations, however, are mainly available for Loobos. The prior value amounts to a significant drop in simulated FAPAR throughout the winter, while the posterior value predicts a much greater longevity of needles with little change in LAI throughout the winter. It appears that the adjustment towards an earlier end of the growing season by changing t_c upwards required a more constant LAI through the year. The posterior value is still smaller than the value reported by *Niinemets and Lukjanova* [2003] for the rate of decline in pine shoots (3.8×10^{-4}), but it is of the same order of magnitude.

6.3. Simulated net primary production

In order to assess to what extent the MERIS FAPAR data have helped to constrain simulations of the net primary production (NPP) of vegetation, we select annual mean NPP at each site as target quantities (i.e. as y in Equ. (4)), including the site for which no FAPAR data were

assimilated. The period chosen for those prognostic simulations is January 2001 to December 2003, which is almost twice as long as the period for which FAPAR data are available. Inferring information for outside the “diagnostic” period is a major strength of the process-based data assimilation technique, as demonstrated before by *Knorr and Kattge* [2005] and *Scholze et al.* [2007].

The computed prior and posterior means and uncertainties of annual NPP are shown in Table 7. Relative change in NPP is again shown as a fraction of the prior uncertainty, which is computed at the optimal parameter point. The lowest NPP is found at the far northern sites, a rather low value also for Loobos, owing to the low value of V_{max}^{25} for PFT 5, and for the semi-arid Maun site, intermediate values for the temperate sites at Hainich and Aardhuis, and high values for the evergreen tropical site at Manaus. Prior uncertainties are considerable for Sodankylä, and moderate for the remaining ones.

The only site where there is a large relative change (absolute value greater than two) in the simulated NPP is Manaus. We suspect that with either larger uncertainties for FAPAR or a more conservative screening algorithm, the posterior NPP would be closer to the prior value. This would also mean much less error reduction for Manaus, which here is shown as 34%. The other sites where we find a considerable uncertainty reduction (by more than 10%) are Aardhuis, a grass site, Hainich forest, and Hainich grass, the latter not included in the data assimilation. This is consistent with a large reduction in uncertainty of T_ϕ and T_r for grasses. Analyzing the model’s Jacobian at the posterior parameter point (see Equ. (4)), we also find that the NPP at Hainich forest has by far the highest sensitivity to Parameter 26 (T_ϕ for PFT 4,5). This explains its relatively large uncertainty (13%) reduction partly as a result of the uncertainty reduction in this parameter (41%).

7. Discussion

The results of this study have shown that with less than two years of FAPAR data, we are able to constrain most of the parameters of a generic model of leaf phenology. With the optimized parameters, the model is able to reproduce the satellite-based FAPAR observations from MERIS within the uncertainty range of the data for three out of the seven sites for which data were assimilated. Agreement was somewhat less satisfactory for the deciduous forest site and for grassland, but agreement still increased substantially after data assimilation. The same was true for a grassland site for which no satellite data were assimilated.

We further expect that the reduction in parameter uncertainty will increase if longer time series are used. We can approximate the effect of doubling the length of the time series on posterior parameter uncertainties by assuming similar values in the first (data) term of Equ. (1) for the additional half of the time series, assuming zero correlation with the uncertainties of the first half of the time series. As a result, the part of the cost function containing the (FAPAR) data is increased by about a factor of two, and the Hessian, which as the cost function can also be written as the sum of a data and parameter driven part, also has its data driven part increased by approximately a factor of two. It is this data driven part of the Hessian that produces the reduction in parameter uncertainty.

The main conclusion of this study is that FAPAR data appear to add significant constraints on those model parameters that control the phenological cycle, namely start, duration and maximum LAI during the growing season. It is also encouraging that some of the posterior parameter values could be shown to agree with independent ecophysiological studies. At the same time, we find that comparatively low FAPAR values at the Manaus site affect several parameters in unexpected ways. Both the Manaus and Hainich sites are known to be dense forests, and the

reason why Hainich has somewhat higher FAPAR than Manaus in the summer may be a result of some residual contamination of FAPAR, in particular by cloud shadows. For future studies, we might consider insuring the functioning of the data assimilation system by increasing the error bars in such cases, once an appropriate indicator has been identified.

Another important result of this study is that simulations of the phenological cycle and NPP seem largely decoupled, causing much less reduction in uncertainty for NPP than for phenological parameters. This was not expected, since FAPAR controls the number of photons available for photosynthesis and net primary production. The result might have been different, had we considered control of leaf phenology from carbon balance [*Kikuzawa, 1995*]. Instead, soil moisture and temperature were chosen as the control state variables of the phenology scheme. We therefore expect that FAPAR would put more constraint on simulated soil moisture than on NPP – this could be the subject of further investigations.

The moderate to small change in the uncertainty of NPP has consequences for the use of light-use efficiency approaches in models and remote-sensing products mentioned in the introduction. For example, the standard MODIS NPP and GPP products [*Zhao et al., 2005*] rely on LAI and FAPAR data derived from MODIS reflectances, climate data, and land use information. This study hints that climate data, land use information combined with the proper parameterization of the algorithm are sufficient sources of information to derive NPP and GPP. The main advantage of using satellite-derived LAI or FAPAR information is that the algorithm used for those products does not require a phenology model – which can certainly be an advantage for purely diagnostic studies.

Another use of remote sensing data was demonstrated by *Sims et al.* [2008], who used remotely sensed land surface temperature in addition to vegetation index data to improve the

climatic information. In this case again, the information is not necessary but additional if a prognostic model is used (i.e one that does not require observations to run). This conclusion also applies to the use of light-use efficiency approaches in global vegetation models [*Field et al.*, 1998; *Zhao et al.*, 2005; *Turner et al.*, 2009].

The alternative to the diagnostic modelling approach used in those products and models is data assimilation in the way presented here, only extended to larger scales. The advantages are that such an approach makes it explicit which information is used for the final end product, that parameters can be used to create a link between observations across space and time, leading to better constraint target quantities (such as NPP or GPP), and that the assimilation system can also be used to make predictions or be run in situations where no observations are available.

We expect that increasing the amount of FAPAR data used in the assimilation will further increase the constraint on NPP. This is because of the specific approach pursued in CCDAS: common parameters that occur at several sites or grid cells are constrained simultaneously. In this way, remotely sensed data from many locations is used to constrain a limited set of parameters. In this case, the data will deliver considerably more constraint on the parameters, and in all likelihood also on NPP.

Therefore, the answer to the question whether FAPAR data products are a useful source of information for carbon fluxes is a qualified yes. They do, provided that the data are used in a way that process knowledge represented by universal parameters is increased. If, however, the data assimilation is carried out at each pixel independently, we would expect only a moderate to small gain for NPP uncertainty.

This is important in the light of a recent study by *Medvigy et al.* [2009] that constrains terrestrial biosphere models with multiple observations, including satellite-derived LAI estimates,

albeit by adjusting a smaller number of dedicated parameters. The results indicate better agreement of simulated carbon fluxes after data constraint. Given the results of that and the present study, we suspect that the main potential of satellite-derived FAPAR data to improve carbon flux simulations lies in the simultaneous use with other constraints, such as eddy covariance and CO₂ data. This would also address the problem of validating the model in areas of high cloud contamination, such as evergreen rainforests.

The results presented here are certainly dependent on the model and on the choice of prior parameters and their uncertainties, so that further application of the approach will be needed to confirm them. However, the study shows that much of the information contained in the FAPAR signal can be reproduced by the model, while the cause for some deviation between modelled and observed FAPAR still needs to be investigated. This is important for the perspective of using FAPAR in a data assimilation mode with a fully prognostic model that is able to simulate the observed signal without the use of observations. It is one of the requirements of data assimilation that the simulated and observed signals are similar and the model is in principle able to reproduce the observed signal [Tarantola, 1987]. We also note that other studies such as the one by Turner *et al.* [2009] have assimilated FAPAR data into an ecosystem model, but no study known to the authors has done so with using a fully prognostic model that is able to simulate leaf phenology from climatic, vegetation type and soils information alone.

8. Conclusions and outlook

We have presented a further development of the Carbon Cycle Data Assimilation in the direction of assimilating remotely sensed information using the same variational approach as previously for the atmospheric carbon dioxide observations. The assimilation is done simultaneously at several sites representing the major biomes of the earth with a common set of parameters. The

algorithm converges rapidly and effectively by highly efficient adjoint model code to a final gradient extremely close to zero.

As such, the study is in principle a global one, only that the number of sample points is kept low. Increasing this number to a global grid is therefore straightforward and is planned to be the subject of a further study. Increasing the number of sites or grid points, as well as the length of the observation period, is also needed to further validate the new generic phenology scheme and to further investigate the causes for some notable deviations between optimal model and observations for grassland and deciduous forest sites.

The study has shown that given the right setup where data are assimilated into a process model that has universal parameters, the method can be used in principle to constrain both parameters of a phenology model, and model-based estimates of carbon fluxes. The approach presented relies on input data of land cover in the form of PFT fractions, soil characteristics in the form of maximum plant available soil moisture, and climate data.

If enough FAPAR data are used to constrain the parameters of the phenology model, it may be possible to constrain in addition the PFT fractions at each site. This would open the perspective of fully automatic, model based land cover mapping complete with characterization of uncertainty ranges. The same applies to the inclusion of maximum plant available soil moisture into the assimilation scheme – such properties cannot be inferred directly from space observations. Applications direct at water resources might require the assimilation of further remotely sensed information, for example from ESA's planned SMOS mission.

Acknowledgments. The authors would like to thank the European Space Agency for financing this project under contract number 20595/07/I-EC, Philippe Goryl and Olivier Colin from ESA/ESRIN, Frascati, for support with the ESA MERIS product, Monica Robustelli and Ioan-

nis Andredakis for help with data processing, Reiner Schnur for provision of meteorological data, Michael Voßbeck for his help with code administration, and two anonymous reviewers and the Associate Editor for helpful comments and suggestions.

References

- Baret, F., et al. (2007), Lai, fapar and fcover cyclopes global products derived from vegetation: Part 1: Principles of the algorithm, *Remote Sens. Env.*, 110, 275–286.
- Botta, A., N. Viovy, P. Ciais, P. Friedlingstein, and P. Monfray (2000), A global prognostic scheme of leaf onset using satellite data, *Global Change Biol.*, 6, 709–725.
- Cao, M., S. Prince, J. Small, and S. Goetz (2004), Remotely sensed interannual variations and trends in terrestrial net primary productivity 1981-2000, *Ecosystems*, 7, 233–242.
- Chen, M., P. Xie, J. Janowiak, and P. Arkin (2002), Global land precipitation: A 50-yr monthly analysis based on gauge observations, *J. Hydrometeorol.*, 3, 249–266.
- Ciais, P., et al. (2005), Europe-wide reduction in primary productivity caused by the heat and drought in 2003, *Nature*, 437, 529–533.
- Deering, D., J. Rouse, R. Haas, and J. Schell (1975), Measuring forage production of grazing units from landsat mss data, in *Proc. 10th Int. Symp. Remote Sensing Environ.*, University of Michigan, Ann Arbor, pp. 1169–1178.
- Enting, I. G. (2002), *Inverse Problems in Atmospheric Constituent Transport*, 392pp pp., Cambridge University Press.
- Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski (1998), Primary production of the biosphere: Integrating terrestrial and oceanic components, *Science*, 281, 237–240.
- Friedlingstein, P., et al. (2006), Climate-Carbon Cycle Feedback Analysis: Results from the C⁴MIP Model Intercomparison, *Journal of Climate*, 19, 3337–3353.
- Giering, R., and T. Kaminski (1998), Recipes for Adjoint Code Construction, *ACM Trans. Math. Software*, 24(4), 437–474, doi:<http://doi.acm.org/10.1145/293686.293695>.

- GLOBALVIEW-CO₂ (2004), Cooperative Atmospheric Data Integration Project - Carbon Dioxide, CD-ROM, NOAA CMDL, Boulder, Colorado, [Also available on Internet via anonymous FTP to ftp.cmdl.noaa.gov, Path: ccg/co2/GLOBALVIEW].
- Gobron, N., B. Pinty, and M. M. Verstraete (1997), Theoretical limits to the estimation of the leaf area index on the basis of visible and near-infrared remote sensing data, *IEEE Trans. Geosci. Remote Sens.*, *35*, 1438–1445.
- Gobron, N., B. Pinty, M. M. Verstraete, and J. Widlowski (2000), Advanced vegetation indices optimized for up coming sensors: Design, performance and applications, *IEEE Trans. Geosci. Remote Sens.*, *38*, 2489–2505.
- Gobron, N., B. Pinty, F. Melin, M. Taberner, and M. M. Verstraete (2002), Sea wide field-of-view sensor (seawifs) – an optimized fapar algorithm – theoretical basis document, *Tech. rep.*, Joint Research Centre, Ispra, Italy.
- Gobron, N., B. Pinty, F. Mélin, M. Taberner, M. M. Verstraete, A. Belward, T. Lavergne, and J.-L. Widlowski (2005), The state of vegetation in europe following the 2003 drought, *Int. J. Remote Sens.*, *26*, 2013–2020.
- Gobron, N., B. Pinty, O. Ausedat, M. Taberner, O. Faber, F. Mlin, T. Lavergne, M. Robustelli, and P. Snoeij (2008), Uncertainty estimates for the FAPAR operational products derived from MERIS – impact of top-of-atmosphere radiance uncertainties and validation with field data, *Remote Sens. Environ.*, *112*, 1871–1883.
- Huete, A. (2007), A soil-adjusted vegetation index (SAVI), *Remote Sens. Environ.*, *25*, 295–309.
- Jarvis, P., and K. McNaughton (1986), Stomatal control of transpiration: scaling up from leaf to region, *Adv. Ecol. Res.*, *15*, 1–49.
- Jones, H. (1983), *Plants and Microclimate*, 323pp pp., Cambridge University Press.

- Jones, P., T. Osborn, K. Briffa, C. Folland, B. Horton, L. Alexander, D. Parker, and N. Rayner (2001), Adjusting for sampling density in grid-box land and ocean surface temperature time series, *J. Geophys. Res.*, *106*, 3371–3380.
- Kaduk, J., and M. Heimann (1996), A prognostic phenology scheme for global terrestrial carbon cycle models, *Clim. Res.*, *6*, 1–19.
- Kaminski, T., M. Heimann, and R. Giering (1999), A coarse grid three dimensional global inverse model of the atmospheric transport, 2, Inversion of the transport of CO₂ in the 1980s, *J. Geophys. Res.*, *104*(D15), 18,555–18,581.
- Kaminski, T., W. Knorr, P. Rayner, and M. Heimann (2002), Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle, *Global Biogeochemical Cycles*, *16*(4), 14–1–14–16.
- Kaminski, T., R. Giering, M. Scholze, P. Rayner, and W. Knorr (2003), An example of an automatic differentiation-based modelling system, in *Computational Science – ICCSA 2003, International Conference Montreal, Canada, May 2003, Proceedings, Part II, Lecture Notes in Computer Science*, vol. 2668, edited by V. Kumar, L. Gavrilova, C. J. K. Tan, and P. L’Ecuyer, pp. 95–104, Springer, Berlin.
- Kikuzawa, K. (1995), Leaf phenology as an optimal strategy for carbon gain in plants, *Can. J. Bot.*, *76*, 158–164.
- Kleidon, A., and M. Heimann (1999), Deep-rooted vegetation, amazonian deforestation, and climate: results from a modelling study, *Glob. Ecol. Biogeogr.*, *8*, 397–405.
- Knorr, W. (2000), Annual and interannual CO₂ exchanges of the terrestrial biosphere: Process-based simulations and uncertainties, *Glob. Ecol. Biogeogr.*, *9*(3), 225–252.

- Knorr, W., and J. Kattge (2005), Inversion of terrestrial biosphere model parameter values against eddy covariance measurements using monte carlo sampling, *Global Change Biology*, *11*, 1333–1351.
- Knorr, W., and J.-P. Schulz (2001), Using satellite data assimilation to infer global soil moisture and vegetation feedback to climate, in *Remote Sensing and Climate Modeling: Synergies and Limitations*, edited by M. Beniston and M. M. Verstraete, pp. 273–306, Kluwer Academic Publisher, Dordrecht and Boston.
- Knorr, W., N. Gobron, M. Scholze, T. Kaminski, and B. Pinty (2005a), Global-Scale Drought Caused Atmospheric CO₂ Increase, *EOS, Transactions*, *86*(18), 178 & 181.
- Knorr, W., I. Prentice, J. House, and E. Holland (2005b), Long-term sensitivity of soil carbon turnover to warming, *Nature*, *433*, 298–301.
- Knorr, W., N. Gobron, M. Scholze, T. Kaminski, R. Schnur, and B. Pinty (2007), The impact of terrestrial biosphere carbon exchanges on the anomalous CO₂ increase in 2002 - 2003, *Geophysical Research Letters*, *34*, doi:10.1029/2006GL029,019, doi: <http://dx.doi.org/10.1029/2006GL029019>.
- Knyazikhin, Y., J. Martonchik, D. Diner, R. Myneni, M. Verstraete, B. Pinty, and N. Gobron (1998a), Estimation of vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from atmosphere corrected misr data, *J. Geophys. Res.*, *103*, 32,239–32,256.
- Knyazikhin, Y., J. V. Martonchik, R. B. Myneni, D. J. Diner, and S. W. Running (1998b), Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from modis and misr data, *J. Geophys. Res.*, *103*, 32,257–32,274.

- Liu, Q., L. Gu, R. E. Dickinson, Y. Tian, L. Zhou, and W. M. Post (2008), Assimilation of satellite reflectance data into a dynamical leaf model to infer seasonally varying leaf areas for climate and carbon models, *J. Geophys. Res.*, *113*, doi:10.1029/2007JD009645.
- Medvigy, D., C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft (2009), Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem demography model version 2, *J. Geophys. Res.*, *114*, doi:10.1029/2008JG000812.
- Monsi, M., and T. Saeki (1953), Über den Lichtfaktor in den Pflanzengesellschaften und seine Bedeutung für die Stoffproduktion, *Jap. J. Botany*, *14*, 22–52.
- Nepstad, D., et al. (1994), The role of deep roots in the hydrological and carbon cycles of amazonian forests and pastures, *Nature*, *372*, 666–669.
- Niinemets, Ü., and A. Lukjanova (2003), Needle longevity, shoot growth and branching frequency in relation to site fertility and within-canopy light conditions in *pinus sylvestris*, *Ann. For. Sci.*, *60*, 195–208.
- Nijssen, B., R. Schnur, and D. Lettenmaier (2001), Retrospective estimation of soil moisture using the VIC land surface model, 1980-1993, *J. Climate*, *14*(8), 1790–1808.
- Oleksyn, J., M. G. Tjoelker, and P. B. Reich (1992), Growth and biomass partitioning of populations of European *pinus sylvestris* L. under simulated 50° and 60° n daylengths: evidence of photoperiodic ecotypes, *New Phytol.*, *120*, 561–574.
- Pinty, B., T. Lavergne, J.-L. Widlowski, N. Gobron, and M. M. Verstraete (2009), On the need to observe vegetation canopies in the near-infrared to estimate visible light absorption, *Remote Sens. Env.*, *113*, doi:10.1016/j.rse.2008.08.017.
- Pinty, B., et al. (2007), Retrieving surface parameters for climate models from MODIS-MISR albedo products, *J. Geophys. Res.*, *112*, doi:http://doi.acm.org/10.1029/2006JD008105.

- Prince, S. D. (1991), A model of regional primary productivity for use with coarse resolution satellite data, *Int. J. Remote Sensing*, *12*, 1313–1330.
- Raddatz, T. J., C. Reick, W. Knorr, J. Kattge, E. Roeckner, R. Schnur, K.-G. Schnitzler, P. Wetzel, and J. Jungclaus (2007), Will the tropical land biosphere dominate the climate carbon cycle feedback during the 21st century?, *Clim. Dyn.*, doi:10.1007/s00382-00007-00247-00388.
- Rayner, P., M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann (2005), Two decades of terrestrial Carbon fluxes from a Carbon Cycle Data Assimilation System (CCDAS), *Global Biogeochemical Cycles*, *19*, doi:10.1029/2004GB002,254.
- Roujean, J., and F. Bréon (1995), Estimating par absorbed by vegetation from bidirectional reflectance measurements, *Remote Sens. Env.*, *51*, 375–348.
- Ruimy, A., G. Dedieu, and B. Saugier (1994), Methodology for the estimation of terrestrial net primary production from remotely sensed data, *J. Geophys. Res.*, *99*, 5263–5283.
- Scholze, M. (2003), Model studies on the response of the terrestrial carbon cycle on climate change and variability, Examensarbeit, Max-Planck-Institut für Meteorologie, Hamburg, Germany.
- Scholze, M., T. Kaminski, P. Rayner, W. Knorr, and R. Giering (2007), Propagating uncertainty through prognostic CCDAS simulations, *J. Geophys. Res.*, *112*, doi:dx.doi.org/10.1029/2007JD008642.
- Sims, D. A., et al. (2008), A new model of gross primary productivity for north american ecosystems based solely on the enhanced vegetation index and land surface temperature from modis, *Remote Sens. Env.*, *112*, 1633–1646.
- Tarantola, A. (1987), *Inverse Problem Theory - Methods for Data Fitting and Model Parameter Estimation*, Elsevier Sci., New York.

- Turner, D. P., W. D. Ritts, S. Wharton, C. Thomas, R. Monson, T. A. Black, and M. Falk (2009), Assessing fpar source and parameter optimization scheme in application of a diagnostic carbon flux model, *Remote Sens. Env.*, *113*, 1529–1539.
- Verstraete, M. M., and B. Pinty (1996), Designing optimal spectral vegetation indices for remote sensing applications, *Trans. Geosci. Remote Sens.*, *34*, 1254–1265.
- Verstraete, M. M., B. Pinty, and R. B. Myneni (1996), Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing, *Remote Sens. Environ.*, *58*, 201–214.
- White, M., P. Thornton, and S. Running (1997), A continental phenology model for monitoring vegetation responses to interannual climatic variability, *Global Biogeochem. Cycles*, *11*, 217–234.
- Wilson, M. F., and A. Henderson-Sellers (1985), A global archive of land cover and soils data for use in general-circulation climate models, *Journal of Climatology*, *5*(2), 119–143.
- Woodward, F. I. (1987), *Climate and Plant Distribution*, Cambridge University Press, Cambridge.
- Zeng, N., H. Qian, C. Rödenbeck, and M. Heimann (2005), Impact of 1998-2002 midlatitude drought and warming on terrestrial ecosystems and the global carbon cycle, *Geophys. Res. Lett.*, *32*, doi:doi:10.1029/2005GL024607.
- Zhao, M., F. A. Heinsch, R. R. Nemani, and S. W. Running (2005), Improvements of the modis terrestrial gross and net primary production global data set, *Remote Sens. Env.*, *95*, 164–176.

Table 1. List of the plant functional types (PFTs) with their corresponding numbers in the global version of BETHY, and the associated phenology.

PFT	Description	Phenology
1	Tropical broadleaf evergreen tree	warm-evergreen
2	Tropical broadleaf deciduous tree	warm-deciduous
4	Temperate deciduous tree	cold-deciduous
5	Evergreen coniferous tree	cold-evergreen
8	Deciduous understorey shrub	cold-deciduous
9	C3 grass	grass
10	C4 grass	grass

Table 2. Summary of the parameters of the phenology model.

Symbol	description	units
$\tilde{\Lambda}$	Maximum leaf-area index	-
T_ϕ	Temperature at leaf onset	$^{\circ}\text{C}$
T_r	Spatial range (1σ) of T_ϕ	$^{\circ}\text{C}$
t_c	Daylength at leaf shedding	hours
t_r	Spatial range (1σ) of t_c	hours
ξ	Initial linear leaf growth	days^{-1}
k_L	Inverse of leaf longevity, τ_L	days^{-1}
τ_W	Length of dry spell before leaf shedding	days

Table 3. Differentiation between phenology types. * denotes a fixed parameter value indicating that this limitation is not active. - shows that the parameter is not active because of the choice of T_ϕ . All other parameter values are prior estimates and are optimized during data assimilation.

PFT	phenology type	growth trigger		deciduousness	
		temperature	day length	temperature	water
		T_ϕ [°C]	t_c [h]	τ_L [days]	τ_W [days]
1	warm-evergreen	$-\infty^*$	0*	-	360
2	warm-deciduous	$-\infty^*$	0*	-	50
4	cold-deciduous	10	10.5	10	0*
5	cold-evergreen	10	10.5	333	0*
8	cold-deciduous	8	10.5	10	0*
9,10	grass	2	0*	10	50

Table 4. Process parameters and their initial and optimized values and prior and posterior uncertainties; rel. change is posterior minus prior divided by prior uncertainty. Top: physiology, bottom: phenology. Units of physiology parameters are V_{max} : $\mu\text{mol}(\text{CO}_2)\text{m}^{-2}\text{s}^{-1}$, k : $\text{mmol}(\text{air})\text{m}^{-2}\text{s}^{-1}$, $a_{\Gamma,T}$: $\mu\text{mol}(\text{CO}_2)\text{mol}(\text{air})^{-1}\text{C}^{-1}$, K_C in $\mu\text{mol}(\text{CO}_2)\text{mol}(\text{air})^{-1}$, K_O in $\text{mol}(\text{O}_2)\text{mol}(\text{air})^{-1}$, activation energies E in J mol^{-1} , others unitless. Uncertainties represent one standard deviation, except for the log-normally distributed parameters denoted by (*), for which relative change in log-space and the analogous difference between mean and upper 67%-tile is given.

Number	PFTs	Parameter	prior value	posterior value	rel. change [%]	prior unc.	posterior unc.	unc. reduction [%]
1	1	V_{max}^{25}	60	44	-133	12	11.1	6
2	2	V_{max}^{25}	90	101.3	63	18	17.5	13
3	4	V_{max}^{25}	35	33.3	-24	7	7	0
4	5	V_{max}^{25}	29	29.2	3	5.8	5.8	0
5	8	V_{max}^{25}	160	160	0	32	32	0
6	9	V_{max}^{25}	42	47.8	69	8.4	7.7	8
7	10	V_{max}^{25}	8	7.4	-38	1.6	1.6	0
8	1	$a_{J,V}$	1.96	1.97	10	0.098	0.098	0
9	2	$a_{J,V}$	1.99	2	10	0.0995	0.0995	0
10	4	$a_{J,V}$	2	2	0	0.1	0.1	0
11	5	$a_{J,V}$	1.79	1.79	0	0.0895	0.0895	0
12	8	$a_{J,V}$	1.66	1.66	0	0.083	0.083	0
13	9	$a_{J,V}$	1.9	1.9	0	0.095	0.095	0
14	10	k^{25}	140	140	0	28	28	0
15	all	E_{R_d}	45,000	45,027	1	2,250	2,250	0
16	all	$E_{V_{max}}$	58,520	56,735	-61	2,926	2,923	0
17	all exc. 10	E_{K_O}	35,948	35,807	-8	1,797	1,797	0
18	all exc. 10	E_{K_C}	59,356	60,611	42	2,967	2,965	0
19	10	E_k	50,967	50,964	0	2,548	2,548	0
20	all exc. 10	α_q	0.28	0.292	86	0.014	0.014	1
21	10	α_i	0.04	0.04	0	0.002	0.002	0
22	all exc. 10	K_C^{25}	460	462	9	23	23	0
23	all exc. 10	K_O^{25}	0.33	0.3303	2	0.0165	0.0165	0
24	all exc. 10	$a_{\Gamma,T}$	1.7	1.66	-47	0.085	0.008	0
25	all	Λ	5	4.2	-320	0.25	0.24	5
26	4,5	T_ϕ	10	9.21	-158	0.5	0.29	41
27	8	T_ϕ	8	8.02	4	0.5	0.5	0
28	9,10	T_ϕ	3	1.92	-36	3	0.23	54
29	1,2,4,5,8	T_r	2	2.04	40	0.1	0.1	1
30	9,10	T_r	2	0.3	-85	2	0.05	50
31	4,5,8	t_c	10.5	13.37	574	0.5	0.31	38
32	4,5,8	t_r	0.5	0.48	-20	0.1	0.1	1
33	all	ξ	0.5	0.37	-130	0.1	0.09	8
34	all exc. 5	k_L	0.1	0.07	-60	0.05	0.012	75
35	5	k_L	3×10^{-3}	1.3×10^{-4}	191	1.5×10^{-3}	8.9×10^{-4}	41
36	1	τ_W (*)	360	1073	269	180	172	63
37	2	τ_W (*)	50	114	203	25	17	65
38	9,10	τ_W (*)	50	23	-192	25	5	48

Table 5. List of sites for assimilation with central coordinates, elevation in m, N-S and E-W extent in km of the rectangular satellite scenes, and n the number of daily data points after spatial averaging. The site in the last row has been included for validation only.

Site	Country	Latitude	Longitude	Elevation	N-S	E-W	n
Sodankylä	Finland	67.3619°N	26.6378°E	180	1.2	1.2	80
Zotino	Russia	60.8008°N	89.2657°E	116	1.2	1.2	101
Aardhuis	Netherlands	52.2381°N	5.8672°E	7	1.2	1.2	91
Loobos	Netherlands	52.1679°N	5.7440°E	25	1.2	1.2	103
Hainich forest site	Germany	51.0793°N	10.4520°E	430	1.2	1.2	106
Manaus	Brazil	2.5892°S	60.1311°W	80	18.0	14.4	146
Maun	Botswana	19.9155°S	23.5605°E	940	3.6	3.6	154
Hainich grass site	Germany	51.0199°N	10.4348°E	302	2.4	1.2	119

Table 6. Site descriptions and site specific parameters. First to third PFT in order of most to least dominant, “fraction” is the associated fraction of surface area in percent and $W_{s,max}$ is maximum plant-available soil moisture in mm. Soil class refers to the brightness of the soil.

Site	Description	PFT			fraction			$W_{s,max}$	soil class
		1st	2nd	3rd	1st	2nd	3rd		
Sodankylä	Boreal evergreen forest	5	4	-	15	10	-	156	medium
Zotino	Boreal mixed forest	5	4	-	20	15	-	105	medium
Aardhuis	C3 grassland	9	4	-	60	20	-	101	medium
Loobos	Temperate pine forest	5	8	9	20	15	5	101	medium
Hainich forest	Temperate deciduous forest	4	9	-	80	10	-	101	dark
Manaus	Tropical rainforest	1	10	-	70	10	-	800	medium
Maun	Tropical savanna	2	10	-	20	10	-	150	medium
Hainich grass	C3 grassland	9	4	-	80	5	-	50.5	dark

Table 7. Mean annual prior and posterior NPP for the period 2000–2003 (inclusive) with uncertainty, change relative to prior uncertainty, and relative uncertainty reduction. Units are $\text{gC m}^{-2} \text{ yr}^{-1}$ or percentage when stated.

Site	prior NPP	post. NPP	rel. change [%]	prior unc.	post. unc.	unc. reduction [%]
Sodankylä	137	151	68	112	98	5
Zotino	201	216	54	28	28	0
Aardhuis	853	842	-7	164	101	38
Loobos	449	424	-40	62	59	5
Hainich forest	689	657	-29	112	98	13
Manaus	1465	964	-196	255	168	34
Maun	350	346	-10	50	46	8
Hainich grass	619	786	0.97	172	89	48

Figure 1. Sketch of the previously used (left hand panel) and new (right hand panel) CCDAS structures: ovals represent input and output data, and grey boxes represent calculation steps. Shown in bold are the state variables of the phenology and hydrology scheme. Diagnostics are quantities of interest such as carbon fluxes computed by CCDAS. Uncert. stands for uncertainty and param. for parameters.

Figure 2. Observed (crosses with uncertainty ranges) and modelled prior (dotted) and posterior (solid line) FAPAR for all eight sites from north (top left) to south. Numbers are root-mean-squared deviation between model and satellite data for the prior (grey) and posterior case (black).

Figure 3. Previous figure continued. The Hainich grass site (bottom right) is shown for validation and not included in the assimilation.