Remote Sensing of Environment 120 (2012) 219–235

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

An Earth Observation Land Data Assimilation System (EO-LDAS)

P. Lewis [a,*], J. Gómez-Dans [a], T. Kaminski [b], J. Settle [c], T. Quaife [d], N. Gobron [e], J. Styles [f], M. Berger [g]

[a] Department of Geography, UCL, and National Centre for Earth Observation, Gower St., London, WC1E 6BT, UK
[b] FastOpt, Lerchenstr. 28a, D-20767 Hamburg, Germany
[c] National Centre for Earth Observation, University of Reading, Reading RG6 6AL, UK
[d] College of Life and Environmental Sciences, University of Exeter and National Centre for Earth Observation, Peter Lanyon Building, Penryn, Cornwall, TR10 9EZ, UK
[e] European Commission, DG Joint Research Centre, Institute for Environment and Sustainability, Global Environment Monitoring Unit, TP 272, via Enrico Fermi 2749, I-21027 Ispra (VA), Italy
[f] Assimila Ltd., 1 Earley Gate, Reading RG6 6AT, UK
[g] ESA ESRIN, Science Strategy, Coordination and Planning Office (EOP-SA), Via Galileo Galilei, Casella Postale 64, 00044 Frascati (RM), Italy

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form 21 December 2011
Accepted 21 December 2011
Available online 18 February 2012

Keywords:
Data assimilation
Vegetation monitoring
Radiative transfer
Sentinel-2
Sentinel-3
Medium to moderate-resolution optical constellations
Leaf Area Index
Chlorophyll

ABSTRACT

Current methods for estimating vegetation parameters are generally sub-optimal in the way they exploit information and do not generally consider uncertainties. We look forward to a future where operational data assimilation schemes improve estimates by tracking land surface processes and exploiting multiple types of observations. Data assimilation schemes seek to combine observations and models in a statistically optimal way taking into account uncertainty in both, but have not yet been much exploited in this area.

The EO-LDAS scheme and prototype, developed under ESA funding, is designed to exploit the anticipated wealth of data that will be available under GMES missions, such as the Sentinel family of satellites, to provide improved mapping of land surface biophysical parameters. This paper describes the EO-LDAS implementation, and explores some of its core functionality. EO-LDAS is a weak constraint variational data assimilation system. The prototype provides a mechanism for constraint based on a prior estimate of the state vector, a linear dynamic model, and Earth Observation data (top-of-canopy reflectance here). The observation operator is a non-linear optical radiative transfer model for a vegetation canopy with a soil lower boundary, operating over the range 400 to 2500 nm. Adjoint codes for all model and operator components are provided in the prototype by automatic differentiation of the computer codes.

In this paper, EO-LDAS is applied to the problem of daily estimation of six of the parameters controlling the radiative transfer operator over the course of a year (>2000 state vector elements). Zero and first order process model constraints are implemented and explored as the dynamic model. The assimilation estimates all state vector elements simultaneously. This is performed in the context of a typical Sentinel-2 MSI operating scenario, using synthetic MSI observations simulated with the observation operator, with uncertainties typical of those achieved by optical sensors supposed for the data.

The experiments consider a baseline state vector estimation case where dynamic constraints are applied, and assess the impact of dynamic constraints on the *a posteriori* uncertainties. The results demonstrate that reductions in uncertainty by a factor of up to two might be obtained by applying the sorts of dynamic constraints used here. The hyperparameter (dynamic model uncertainty) required to control the assimilation are estimated by a cross-validation exercise. The result of the assimilation is seen to be robust to missing observations with quite large data gaps.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

1.1. Background

One of the primary goals of Earth Observation (EO) is to provide objective and reliable information on the current and (particularly within the satellite EO era) historical state and dynamics of the Earth environment. A major component of this that has been a significant focus of research efforts on monitoring terrestrial vegetation, but EO data are usually of a radiometric nature and do not give direct estimates of the properties of the Earth land surface that we wish to map. Some level of *inference* is therefore needed.

Early studies in terrestrial vegetation monitoring from EO (Asrar et al., 1985; Choudhury, 1987; Richardson & Wiegand, 1977; Tucker, 1979) found that simple transformations of multispectral measurements at red and near infrared wavelengths gave a signal that was responsive to the relative amount of green biomass and that could be used to track vegetation dynamics (Behrenfeld et al., 2001; Goward et al., 1985; Nemani et al., 2003). The attractions of such 'Vegetation Indices' (VIs) are obvious: they are visually impressive as spatial

* Corresponding author. Tel.: +44 20 7679 0585.
E-mail address: plewis@geog.ucl.ac.uk (P. Lewis).

0034-4257/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2011.12.027

and temporal datasets; they are simple to produce and provide a single quantity to interpret; they compensate for some of the extraneous factors that can otherwise complicate lower level EO signals; and they can often provide effective information for time series analyses, where the timing, rather than the magnitude of events is of importance (e.g. vegetation phenology). Further, such indices can be directly targeted at particular functional or physical vegetation properties, such as the fraction of absorbed photosynthetically active radiation (fAPAR) or Leaf Area Index (LAI), by design (Gobron et al., 2002; 2010) or empirically (Rochdi & Fernandes, 2010). In the former case a calibration is achieved using a set of radiative transfer model runs over a range of conditions (Gobron et al., 2000). In the latter, extensive ground-based measurements must be made (J. M. Chen et al., 2002) and the form of the relationship with a particular VI assumed. Such efforts are fast to process and often effective, especially for near-real-time survey. They have a range of known failings (Baret & Guyot, 1991), but some of these, such as dependence on the angular conditions of data acquisition can be reduced by treating the data to normalise for such effects (e.g. Rochdi & Fernandes, 2010). Ultimately though, however much care is taken to treat such effects, methods assuming such fixed mappings from VIs with 'statistical' models are open to many criticisms, some of the more significant of which could be considered: (i) they fail to make full use of the information content of the observational data; (ii) they (often) fail to make use of our understanding of the physics of the situation; (iii) they need recalibration if conditions change (e.g. sensor band pass functions or scale of observation); (iv) they tend not to treat uncertainty in the mapped product in any rigorous way (mostly, they fail to consider this at all).

An alternative stratagem has been to build mathematical models of the physics of radiation interactions with vegetation canopies and the intervening atmosphere, phrased as functions of 'control' variables (polarisation, wavebands, viewing and illumination angles etc.) and (bio)physical parameters or 'state variables' (LAI, leaf chlorophyll concentration etc. for the canopy, and aerosol optical depth, ozone concentration etc. for the atmosphere), and to use these to interpret the satellite signal. We may call these radiative transfer (RT) models. To tie in with discussions below and to provide consistency with the data assimilation literature, such models are called here 'observation operators' (denoted$H(x)$) in that they map from the state variable vector x to the EO signal (as a vector) R for a given set of control variables, so the modelled signal vector $R = H(x)$. The 'remote sensing inverse problem' then is to obtain an estimate of some function of x, $F(x)$ from measurements$R$. How this may be achieved is discussed in more detail below.

Much effort has been devoted to producing information from EO data about specific biophysical quantities that are relevant to science and society. A major focus of this has been to attempt to provide estimates of (green) LAI. Garrigues et al. (2008) consider four representative EO-derived global LAI products, with core spatial resolutions of 1 km or coarser, that use what might be considered state of the art methods for multi-year dataset generation. The reader is referred to that paper for detailed information on the products, a product intercomparison and validation against independent ground measurements. The temporal resolution of the products varies from 8 days to 1 month. Three of the products (ECOCLIMAP, GLOBCARBON (V1), and CCRS) are derived from assumed VI relationships with LAI. A fourth (MODIS (C4)) uses such a relationship for a backup algorithm. Three of the products (GLOBCARBON, CYCLOPES (V3.1) and MODIS) make use of RT models in attempting to estimate the LAI. In the case of GLOBCARBON the RT model is used to calibrate the VI-LAI relationship. For MODIS a look-up table derived from the RT model is used to map red and near infrared (NIR) bidirectional reflectance data to LAI, and for CYCLOPES a neural network derived from an RT model is used for the mapping from red, NIR and shortwave infrared (SWIR) portions of the electromagnetic spectrum. A feature of these uses of RT models is that they can map many channels of input data to one (or many) outputs. The one-to-one mapping used in VI design and/or calibration is then just the simplest case of this more general RT approach.

A major new effort in satellite data provision is the GMES (Global Monitoring for Environment and Security; www.gmes.info) programme (Council of the European Union, 2010). It is an EU initiative set up to provide timely information on key environmental variables for policy makers and public authorities, and is intended to be a major EU contribution to understanding and managing climate change. Six thematic areas are being developed: marine, land, atmosphere, emergency and security and climate change. The land monitoring service is provided via the GEOLAND2 project (www.gmes-geoland.info), which oversees the generation of products derived from satellite data, providing information on a wide range of variables including LAI. GMES is a European contribution to GEOSS, the Global Earth Observing System of Systems (European Commission, n.d.). The Sentinels are a series of satellites being developed by the European Space Agency that are specifically designed to address the space observation requirements of GMES. There are five Sentinel missions, each of which will consist of a pair of satellites (for details see Aschbacher and Pérez (2010) and dedicated Sentinel mission papers, all this RSE issue). This paper is primarily concerned with methods for the retrieval of biophysical parameters of terrestrial ecosystems, including LAI, from instruments at arbitrary spatial resolutions, sun-sensor geometries and optical wavelengths. Consequently the techniques described here are directly relevant to Sentinels 2 and 3 missions. Sentinel 2 has a medium resolution multispectral imager (MSI) in the optical domain with 4 bands at a 10-m resolution, 6 bands at 20 m and 3 bands at 60 m. These 13 spectral channels (Table 1) are distributed in the visible and near infrared and shortwave infrared regions. The Ocean Land Color Instrument (OLCI) instrument on board the Sentinel 3 platform is a coarser (circa 500 m) resolution instrument, similar to MERIS that is designed for global monitoring applications. In principle the system described in this paper could also be extended to other wavelength domains and consequently be used to integrate data from the entire suite of EO missions.

An additional context for this paper is the growing interest in the application of wider constellations of satellites for environmental and disaster monitoring. A manifestation of this is the NASA A-train (NASA, n.d.), which is a formation of complementary satellites and sensors taking observations at close to the same time. Other examples include relatively low cost satellites and instruments with a suite of similar instruments flying in formation to provide global daily viewing opportunities at moderate resolution (10–30 m), for example the Disaster Monitoring Constellation (DMC) (DMCII, 2010). The concept can potentially be applied to more heterogeneous systems, such as the 'virtual constellation' for Land Surface Imaging (LSI) concept promoted by the Committee on Earth Observation Satellites (CEOS) to optimise benefits from land remote sensing systems (CEOS,

Table 1
Spatial resolution, central wavelength and bandwidths for Sentinel-2 MSI (ESA, 2010).

#	Spatial resolution/m	Wavelength/nm	Bandwidth/nm
1	60	443	20
2	10	490	65
3	10	560	35
4	10	665	30
5	20	705	15
6	20	740	15
7	20	783	20
8	10	842	115
8a	20	865	20
9	60	945	20
10	60	1375	30
11	20	1610	90
12	20	2190	180

2010). There are clear benefits for monitoring frequency if data from a wider range of sensors are available, but the more heterogeneous the set of sensors (in terms of spatial resolution and wavelength domains), the more important it is to formalize appropriate methods to optimally merge information from these sources.

1.2. Optimal estimation

The remote sensing inverse problem described above can be phrased as an optimal estimation problem, requiring an estimate of a distribution around the minimum of some function of an observation residual vector, such as an χ^2-norm. Our assimilation system is based on the joint inversion approach of (Tarantola, 2005) and is most conveniently formulated in what is often called a Bayesian context (Enting, 2002), which means that each piece of information (including any prior information on the state variables) is represented by a probability density function (PDF). Combining this information yields an *a posteriori* PDF for the parameters, which is the result/solution of the assimilation problem. If all of these PDFs are Gaussian and the models involved not too non-linear (potentially after a transformation) then the posterior parameter PDF can also be approximated by a Gaussian:

$$\rho(x) = \exp(-J(x))$$

which is the maximum likelihood estimate of the state variables x, thus the minimum of a cost function which takes the form:

$$J(x) = \sum_i J_i(x) \quad (1)$$

where $J_i(x)$ is a cost function expressing a constraint i, a member of some set of constraints.

Much of the earlier literature on estimating x for vegetation monitoring from a physical basis concentrated on exploring options in numerical minimisation approaches (see e.g. the review by (Kimes et al., 2000)) based almost entirely on using a single cost function $J_{obs}(x)$ expressing a mismatch between EO data and the prediction of an observation operator $H(x)$ (a radiative transfer model). The optimisation methods explored include, but are not limited to, downhill simplex (Privette et al., 1994), gradient methods (Gill et al., 1981; Liang & Strahler, 2002), neural networks, look-up tables and genetic algorithms (GA) (Combal et al., 2003; Myneni et al., 1995; Weiss et al., 2000). Although appropriate optimisation strategies and computer implementations have been around for some time that make use of the gradient of J_{obs} with respect to$x$, J'_{obs} in locating the minimum, they have not been widely used in terrestrial EO monitoring. This has been primarily because of the perceived computational cost and numerical issues if finite difference methods are used to estimate J'_{obs}, and more particularly because it is no trivial job to differentiate radiative transfer models. The advent of automatic differentiation (AD) methods and tools such as TAF (e.g. Giering & Kaminski, 1998; Lavergne et al., 2007) or TAPENADE (e.g. Qin et al., 2007) means that calculating J'_{obs} for radiative transfer or other models is now quite feasible at computational costs not greatly dissimilar to the calculation of J_{obs}. The approach has first been applied to rather simple RT models such as RPV (Lavergne et al., 2007) and a two-stream model (Clerici et al., 2010), but this is equally appropriate for more complex models as we show here. The ability to make rapid, exact calculation of the gradient vector not only widens the choice of algorithms that might be used to minimise the cost function, but also provides a route for potentially faster state vector estimation, and perhaps most importantly allows larger dimensioned problems to be tackled. Qin et al. (2008) were perhaps the first to apply AD to more complex RT models (MCRM of Kuusk (1995)) using a combination of GA and a cost function-based method using J'_{obs} in the region of a trust region derived from the GA. In this case 7 members of the (dimension 14) state vector are estimated, but only at a single point in time. Results are not shown for parameters other than LAI, and no detailed consideration of uncertainty is included, but the ability to use AD in such scenarios is clearly demonstrated.

Data producers and users generally have little influence over control variables to the estimation problem, as satellite sensors and missions are usually designed to serve (or are used to serve) multiple purposes, and involve compromises in sensor design and orbits. Any one sensor (and the resultant set of control variables) then will tend to be sub-optimal for a task as specific as vegetation monitoring. Inevitably this results in individual EO data sources having information content that is too low to provide accurate retrievals of the entire state vector space. Some parameters may never be completely retrievable on the basis of observation alone, especially where there is equifinality between two or more parameters over the domain of the observed data, that is, when the same observed model state can be reached by different combinations of state variables. See for example Beven (2006) for an overview of this issue or Lewis and Disney (2007) for an attempt at explaining mechanisms impacting this in canopy radiative transfer. The core of the issue is that the observations only refer to a subspace of the unknown state variable space. In this case, no information on some directions in state space can be gained from the observations, and their values will have to be constrained using for example, prior information. Such problems are described as being ill-posed. As an example, consider the often-desired goal of tracking the temporal evolution of some parameter of interest such as LAI, to provide information on phenology. Inverting a model on a daily basis where there may only be a small number of observations, or none at all, is typically not possible as the observations do not have enough information to constrain all of the state vectors of typical radiative transfer models. This has been solved implicitly in the production of many current EO data products by assuming the model parameters to be constant over some time interval, and many of the ancillary parameters such as those governing leaf and soil properties are simply assumed known (and fixed as is the case when using VIs). Assumptions such as temporal invariance or knowledge of ancillary variables are pragmatic responses to the remote sensing problem being ill-posed, but it is better if possible to seek less *ad hoc* methods for constraining our estimate, especially if we wish to estimate uncertainty in the product.

A mechanism that provides scope for dealing with such problems is the suite of tools that are collectively referred to as 'Data Assimilation' (DA). There is no strict definition as to what constitutes DA but it is taken here to mean the statistically optimal merging of data and models. Optimality, in this sense, implies the need to take into account uncertainties in all parts of the system.

1.3. Data assimilation

Data assimilation can be seen as mechanism for combining models and data. The defining feature of DA, at least by the definition provided in this paper, is that it enables the use of additional assumptions to make parameter estimation viable in situations that exhibit ill-posedness. In essence, we have a mechanism through Eq. (1) to combine multiple constraints. An example of this that has long been used either explicitly or implicitly in the inference of land surface parameters from EO is constraint via *a priori* estimates of parameter values or ranges (or more generally, distributions). What DA specifically brings to bear on the problem is a dynamic model of parameter evolution in space and/or time.

Early examples of data assimilation systems are those used to improve short-range weather predictions from meteorological models (Ghil & Malanotte-Rizzoli, 1991). In these systems the number of state variables is typically huge, often greater than 10^6, because of the large number of interconnected sub-domains used to represent

the atmosphere in a 3D grid. The number of observations available is typically several orders of magnitude less than this, and in consequence the problem is ill-posed. However, including a constraint that the final solution should not diverge too far from an *a priori* estimate (typically supplied by a previous model run) tends to result in a tractable solution. The schemes used for these problems are referred to as 'variational', being based in the field of mathematics dealing with the calculus of variations, and are closely related to the DA system described in this paper. A 'strong constraint' variational DA system assumes that the underlying process model prescribing the state vector evolution is correct (i.e. there is a model trajectory that matches the observations). In this case it is generally only the initial state of the system that is estimated by the DA procedure, but this approach can also be used to calibrate models (i.e., to optimise estimates of model process parameters) (Knorr et al., 2010). If the state vector is allowed to deviate from the model predictions then this is referred to as a 'weak constraint' DA system (Zupanski, 1997). It is this latter type that is used here and is discussed more completely in later sections.

We note that these systems have been exploited to estimate LAI from MODIS data by making use of a coupled phenology temporal trajectory model with a radiative transfer model (Xiao et al., 2009, 2011). MODIS LAI is assimilated into a crop model using a variational technique in Fang et al. (2008a, 2008b). The variational approach is shown to help in retrieving surface fluxes in Olioso et al. (2005) and Qin et al. (2007), and has found wide application in the hydrological literature (see for example McLaughlin, 2002)).

Another related set of techniques in the DA canon may be called sequential methods. The most widely-known and widely used example of these is the Kalman Filter and its variants. Sequential methods generally only consider observations at a single time step and adjust the model state vector at that time by an amount proportional to the differences between the observations and the predictions of those observations using that model state. Using a variant of the Kalman Filter, known as the Ensemble Kalman Filter (Evensen, 2003), Quaife et al. (2008) demonstrated the assimilation of satellite reflectance data into a simple ecosystem model using an RT model as observation operator. Other efforts have used these techniques to assimilate e.g. snow data (Slater & Clark, 2009) or MODIS-derived LAI into a phenology model (Stöckli et al., 2008). The related technique of particle filtering has been used to assimilate microwave temperature in order to infer soil moisture dynamics in (Qin et al., 2009).

The Earth Observation Land Data Assimilation System (EO-LDAS) study funded by ESA aims at supporting the generation of a generic land data assimilation system by using the full information content provided by observations from satellite constellations. Such a system, in eventual operational form, is intended primarily to improve the quality and consistency of land surface products generated from multi-sensor EO data. The project is focussed on developing a generic scheme and software prototyping for use with medium to moderate spatial resolution (in the range 10 m–500 m) optical data. The principal design concept is to allow integration of data from different satellites observing the surface of the earth at different sun-sensor geometries, wavebands and spatial scales, such as that supplied by Sentinels 2 and 3, in a physically consistent manner, and to provide information on the state of the surfaces with well-quantified estimates of uncertainty. It also demonstrates the idea that predictions based on data from one sensor can be made from a DA system driven by observations from another, a concept that could potentially be used to aid vicarious sensor calibration.

2. The EO-LDAS prototype

2.1. The EO-LDAS scheme

The EO-LDAS prototype is an initial version of the scheme, designed to carry out a core set of DA functions. In particular, in the *scheme*, it performs an atmospheric correction of images to top-of-canopy reflectance, retrieves canopy state variables using surface reflectance data and a constraint model and simulates top-of-atmosphere radiance or reflectance for a given surface and atmosphere. This preserves the essential features of a more comprehensive system (incorporating a fuller coupling between the surface and atmosphere), while allowing development and further study of the most important elements—the observation operators and the assimilation techniques.

To simplify the prototype, we have assumed a large length scale for variations in atmospheric scattering properties, and a very short length scale for surface variability. With these assumptions, we can correct an image (or sub-image) with a single set of atmospheric state variables, use reflectance data in a multi-temporal assimilation on a cell-by-cell-basis, and simulate a top-of-atmosphere radiance field using the same atmosphere for each of a set of model grid cells. This process can be iterated to achieve the surface-atmosphere coupling. To relax either constraint would mean we have to deal with the inversion of a coupled surface-atmosphere problem over a large number of cells, which would require considerable computing resources, both in terms of memory (for the covariance structures involved) and the time needed to carry out the actual inversion, without necessarily improving our ability to monitor the land surface. A tutorial guide explaining the functionality and use of the prototype system is available online.[1]

The DA system can be considered to have two main components: (i) a set of constraints, expressed via Eq. (1); (ii) an assimilation algorithm, i.e. a way to apply the constraints to achieve the optimal estimate of the state vector. The set of constraints in EO-LDAS involves: (i) an observational constraint $J_{obs}(x)$, requiring data (from EO or ground measurements) and a model for translating from state space to observation space (the observation operator); (ii) a dynamic model constraint $J_{model}(x)$, conditioning the temporal (and/or spatial) evolution of the state vector; (iii) physical or empirical bounds and/or distribution constraints $J_{prior}(x)$ to the state vector elements. Thus, in EO-LDAS, Eq. (1) becomes:

$$J(x) = J_{obs}(x) + J_{prior}(x) + J_{model}(x)$$

Each of these constraints has associated with it an error model. In the following sections, we describe the set of constraints and the DA algorithm. We stress that in the text below, we use the symbol x to refer to the set of state variables that we wish to estimate. In EO-LDAS this essentially means a representation of the state at each sample points in time (and/or space) that we consider. So, for example if we were trying to estimate Leaf Area Index and leaf Chlorophyll content at one location for every day of the year, we would have a state vector with 365×2 elements. In addition, EO-LDAS has the capacity to augment this state vector with 'static' state representations (some term affecting one or more of the constraints that we wish to be considered constant in space/time).

2.2. Observational constraint

Given the EO context of this system, at least one of these constraints should be based on observations. The cost function is generally weighted for observation and observation operator uncertainty and correlation (assumed in EO-LDAS Gaussian and described by its covariance matrix, C_{obs}):

$$J_{obs}(x) = \frac{1}{2}(R - H(x))^T C_{obs}^{-1}(R - H(x)) \quad (2)$$

where ^{T} denotes the transpose operator. This is the penalisation associated with differences between the predicted and observed

[1] http://www.geog.ucl.ac.uk/~plewis/eoldas/.

reflectance values. The covariance matrix C_{obs} describes the uncertainty in the observations (and also formally, in the observation operator). As noted, the purpose of the observation operator $H(x)$is to translate information from the state space to that of the observations, and is in practice a radiative transfer model. For ease of implementation (mainly involving spectral sampling issues), when different sensor types are used in EO-LDAS, a set of $J_{obs}(x)$ terms is developed, with one for each sensor type.

There have been many attempts to create observation operators $H(x)$, varying in complexity, accuracy and computational cost. Goel (1988) provides a review of most of the concepts for radiative transfer model developed for reflectance from vegetation canopies at optical wavelengths (see also Goel and Thompson (2000), with Tha Paw and others (1992) covering related materials for thermal emitted radiation and Fung and Chen (2010) for the microwave domain. Some updates and model intercomparisons are provided by Sobrino et al. (2005) (thermal) and Widlowski et al. (2007) (optical). The focus in this paper and in the prototype EO-LDAS is on the use of optical sensor data, but the approach outlined here is easily adapted for use in other wavelength domains.

In a similar way, atmospheric properties, such as aerosol optical depth and water vapour content, need to be accounted to obtain accurate estimates of surface properties. This can be achieved by coupling the surface model with an atmospheric model, and solving for both the surface and atmosphere parameters simultaneously (Verhoef & Bach, 2003). Some (probably most) approaches to surface interpretation use surface reflectance that has already been 'corrected' for atmospheric effects (Vermote et al., 2002), but a full decoupling of the problem, at optical wavelengths at least, cannot be achieved without knowledge of the surface Bidirectional Reflectance Distribution Function (BRDF) (Lyapustin and Knyazikhin, 2001; Lyapustin et al., 2006) (or at least a normalised form of this) (Vermote et al., 2002).

The observation operator we use in this paper is developed from the original semi-discrete model of Gobron et al. (1997). It has a state vector describing canopy architecture and three spectral terms, although these are all defined as functions of other parameters as described below (Table 2). The soil reflectance is assumed Lambertian in the model, although it could be adapted to incorporate a soil directional reflectance model. As stated here then, the (canopy-soil) model estimates the directional reflectance factor at a set of viewing and illumination angles for a given narrow waveband. Since the model must be capable of predicting the reflectance at arbitrary (solar reflective) wavelengths, spectral models are incorporated in the code to predict the soil (Lambertian) reflectance and leaf (bi-Lambertian) reflectance and transmittance. Since model derivatives are required, we use for simplicity here: (i) the linear soil reflectance model of Price (1990); and (ii) an approximation to the PROSPECT leaf reflectance/transmittance model of Féret et al. (2008), being a minor modification of the model of Jacquemoud and Baret (1990). The approximation was developed for possible processing speed enhancements, but is identical in form to PROSPECT if the parameter N (Table 2) is 1, and very close to the original model over the range of N 0.8 to 2.5.

The soil spectral model of Price (1990) characterises a given soil at field capacity as a linear combination of Empirical Orthogonal Functions (EOFs) based on a database of moist (field capacity) soil spectra. Four EOFs are found to account for 99.6% of the cumulative variance of all the soils considered, so, as is usual, we use up to four terms in this implementation. Parameter ranges in Table 2 come from (Price, 1990), Figs. 11–13.

The leaf angle distribution is categorised in the model of Gobron et al. (1997) and so not set by this assimilation procedure (i.e. it must be pre-defined or the different categories assessed separately: this could ultimately be improved using a continuous description). The assimilation scheme can provide estimates of the remaining (12) state variables for each time period modelled. Following Weiss et al. (2000) we apply approximate linearization functions to some of the terms (Table 3). The reasons this is appropriate here are: (i) they better condition the problem for optimisation; (ii) the assumptions of Gaussian distributions of errors are more appropriate in this case.

Differentiated versions of the observational cost are required to enable the use of efficient gradient descent minimisation routines, so we can benefit from access to $J'_{obs}(x)$, the derivative of $J_{obs}(x)$ with respect to x. This is:

$$J'_{obs}(x) = -H'(x)^T C_{obs}^{-1}(R-H(x)) \quad (3)$$

where $H'(x)$ is the derivative of $H(x)$ with respect to x. An adjoint code of the cost function for the semi-discrete model, i.e. code for direct calculation of $J'_{obs}(x)$ that avoids the need for explicit calculation and storage of $H'(x)$, was generated from the source code of the model by the automatic differentiation tool TAF (Giering & Kaminski, 1998). The adjoint code implements the chain rule of differentiation in the so-called reverse mode. It provides the gradient information that is accurate up to machine precision at a computational cost that is not greatly dependent of the length of the gradient vector and well below that of the multiple runs of the semi-discrete model that would be required for a finite difference estimate.

We obtain an estimate of the posterior uncertainty through consideration of the curvature at the global minimum in state space. This is provided by the inverse of the sum of the constraint Hessians, the Hessian for this constraint being $J''_{obs}(x)$:

$$J''_{obs}(x) = H'(x)^T C_{obs}^{-1} H'(x) - H''(x)^T C_{obs}^{-1}(R-H(x)) \quad (4)$$

Although it should be possible to develop a Hessian code in much the same way as done for the first derivative, that has not yet been done within EO-LDAS, so a linear approximation to the Hessian is achieved, using finite differences. As we will see below, the algorithm used to perform the optimisation is iterative, but the potentially high

Table 2
Summary of observation operator state variables.

#	Name	Symbol	Units	Default value	Lower limit	Upper limit
1	Leaf area index	LAI	none	0.01	0.01	5.4
2	Canopy height	xh	m	5	1.0	5
3	Leaf radius	xr	m	0.01	0.001	0.1
4	Chlorophyll a, b	C_{ab}	μg cm^{-2}	40	0	200
5	Carotenoids	C_{ar}	μg cm^{-2}	0	0	200
6	Leaf water	C_w	cm^{-1}	0.01	0.00001	0.04
7	Dry matter	C_{dm}	G cm^{-2}	0.01	0.00001	0.02
8	Leaf layers	N	None	1.0	1.0	2.5
9	Soil PC 1	s_1	None	0.2	0.05	0.4
10	Soil PC 2	s_1	None	0	−0.1	0.1
11	Soil PC 3	s_1	None	0	−0.05	0.05
12	Soil PC 4	s_1	None	0	−0.03	0.03
13	Leaf angle distribution (categorised)	g	None	uniform	1. Planophile 2. Erectophile 3. Plagiophile 4. Extremophile 5. Uniform	n/a

Table 3
Transformations applied to approximate linearise state variable response.

#	Transformed symbol	Transformation
1	TLAI	$\exp(-LAI/2.0)$
4	TC_{ab}	$\exp(-C_{ab}/100)$
5	TC_{ar}	$\exp(-C_{ar}/100)$
6	TC_w	$\exp(-50\ C_w)$
7	TC_{dm}	$\exp(-100\ C_{dm})$

cost of using finite differences for the Hessian is unimportant in this sense, as it only has a role in estimating the posterior uncertainties.

2.3. Process model constraint

The EO-LDAS prototype is designed to allow the user to interface their own constraints, so long as they provide code to calculate the cost function and its first and second-order derivatives. This allows a mechanism whereby (bio)physical process models can be used to constrain the solution and or estimates of the variables controlling those models can be developed. The focus of the prototype software and that of this paper are on understanding how to use DA concepts to improve estimates of biophysical variables from EO data, rather than to test specific process models however. For this reason, we have currently only implemented a linear process model in the system:

$$M(x) = Ax + b \tag{5}$$

where A and b are a matrix and vector respectively. One advantage of designing the prototype system in this manner is that it provides a flexible framework for changing the underlying model. Unlike in a sequential system, this formulation directly allows for any model state vector element to be linked to any other, since x here contains the state representation at all sample times (spaces), so different time/space scales can be readily incorporated. The cost function associated with this process model then is:

$$\begin{aligned} J_{model}(x) &= \frac{1}{2}(x-M(x))^T C_{model}^{-1}(x-M(x)) \\ &= \frac{1}{2}((I-A)x-b)^T C_{model}^{-1}((I-A)x-b) \end{aligned} \tag{6}$$

where I is the identity operator. $J_{model}(x)$ is the cost incurred by departure of the model state from that predicted by an underlying process model. An interpretation of A is as the model derivative. The model uncertainty matrix C_{model} therefore expresses the uncertainty in this derivative, including any inherent uncertainty in the process model. It such a case, it might often be pragmatic to specify only diagonal terms in C_{model} as further details of model structure are often difficult to obtain. In any case, we can see that EO-LDAS could be interfaced to a process model such as the Carbon Flux model DALEC used by Quaife et al. (2008) or any other for which the derivative might be obtained (e.g. using AD) by augmenting the state vector x by any terms that we might wish to drive the model.

Whilst the EO-LDAS scheme allows for linking to 'biophysical' or other process models, that is not the main focus of the prototype. Indeed, there are many cases, for instance when conducting a comparison of information derived from EO data and some biophysical model trajectory, when it may be undesirable to directly incorporate a detailed process model. Further, and perhaps more importantly, a fundamental requirement of the EO-LDAS system is that the state vector, x, contains at least the parameters of the observation operator $H(x)$ for every point in time (and/or space), and many of these may not be provided by a biophysical process model designed, for example, to estimate total Carbon fluxes. We should see the matrix A (and if needed, the vector b) then as a much more general interface to 'process modelling' within an optimal estimation environment.

We can for example consider the benefits of approaches such as Twomey-Tikhonov regularisation or variations around this theme (Rodgers, 2000). Examples of this that we explore further below are first and second-order difference constraints. In essence these improve the conditioning of the inverse problem by smoothing or regularising the solution, which comes about because they constrain derivatives (first or second order here) to be zero. In a weak constraint DA system such as that used here, the model is not strictly enforced (this would be clearly undesirable in these derivative constraints) but rather the degree of smoothness in the outcome is traded off against the other factors in $J(x)$ through the model uncertainty matrix. In other words, the cost function will penalise temporal trajectories of parameters that are not flat, but this is 'balanced' with a goodness of fit to the observations and departure from the prior estimate. In practice this constrains the solution toward a smooth evolution by minimising the high frequency components of the temporal parameter trajectory. A similar approach has been taken by Quaife and Lewis (2010) for linear observation operators. Viewing this form of solution as a combination of state variable estimation and filtering, we note that the filter characteristics are controlled by the nature of matrices A and C_{model}, the former controlling the cut-off frequency of the filter and the latter, if simply diagonal, controlling the degree of dampening of the unwanted high frequencies. In this context, we can consider b a bias term, which we set to zero. In this case:

$$J_{model}(x) = \frac{1}{2}(Dx)^T C_{model}^{-1}(Dx)$$

where $D=(I-A)$. The derivatives of this are: $J'_{model}(x) = D^T C_{model}^{-1} Dx$ and $J''_{model}(x) = D^T C_{model}^{-1} D$. To achieve Twomey–Tikhonov regularisation then, which we view as an empirical process model, D here becomes simply a (N^{th} order) differential operator (Quaife & Lewis, 2010).

In many situations, we must assume the uncertainty in this empirical constraint unknown. The minimum error model then is a constant value for which we can use a scalar term γ:

$$J_{model}(x) = \frac{\gamma^2}{2} x^T \left(D^T D\right) x \tag{7}$$

We can interpret γ as a 'smoothness term (or γ^{-1} as a roughness term) that controls the weighting of the derivative (model) constraint with respect to the other constraints. It is worthwhile at this point trying to relate this back to the discussions on process models. This is most readily achieved by considering a first order derivative constraint. If applied at lag 1 day for a temporal constraint, we can interpret this as an expectation that the state vector tomorrow will be the same as today (i.e. the derivative is zero). If we want to relate this to equivalent sequential methods, we can say that this is a zero-order process model. The term γ^{-1} then can be interpreted as uncertainty (phrased as standard deviation) in this model, or alternatively as the growth in uncertainty over a one day period. Similar interpretations apply for other derivative constraints: a second-order derivative constraint is equivalent to a first order process model. Eq. (7) then is a viable empirical process model constraint, but we have yet to tackle the fact that the smoothness γ is unknown. We also note that if we use a scalar for γ, we are assuming the same smoothness for all state variables at all times (places).

An option that arises with dynamic models (where we are making connections between elements of the state vector at different times (places) is what to do about boundary conditions. Even with a simple differential model this needs consideration in forming the D matrix. Among the various options, especially when dealing with annual or multi-annual datasets, an attractive one is to assume periodic boundary conditions, and that is done here. This means that in calculating D at the end of the year (edge of the matrix) we perform the digital differential with state elements from the beginning of the year.

It is generally found (e.g. Twomey (2002)) that quite a broad range of model uncertainty (smoothness) estimates can provide an acceptable solution, so we do not expect the results to be overly-sensitive to the choice of this 'hyperparameter'. We could make a rough guess at the model uncertainty, but that is likely to be unsatisfactory in the general case. If we underestimate it by too much, we can over-dampen most of the state vector. Equally, if we greatly

over-estimate the model uncertainty, the impact of the temporal constraints is minimal: in the extreme, an infinite model uncertainty (zero smoothness) leads to a solution without model constraint. Whilst there are several strategies that can be employed to estimate the model uncertainties (hyperparameters), perhaps the most fruitful in the context of EO-LDAS is running a cross-validation exercise. The idea is that an independent dataset is used to test the robustness of the solution for a particular value of the hyperparameters. An optimal estimate of the hyperparameters (or distribution thereof) can be obtained by minimising a cost function with the independent observations. This can be achieved with a subset of observations to test a solution obtained from the rest of the dataset, a strategy that when repeated over different subsets becomes known as generalised cross validation (Eilers, 2003; Lubansky et al., 2006; Wahba, 1990). Alternatively, we might use data from an independent sensor, although accurate absolute calibration between the sensors is needed for that.

2.4. Prior constraint

An additional constraint mechanism is implemented in EO-LDAS, that we term a prior constraint. Its role, via the cost function $J_{prior}(x)$ is to impose a penalty for deviation from some previously defined state, x_{prior}:

$$J_{prior}(x) = \frac{1}{2}\left(x - x_{prior}\right)^T C_{prior}^{-1}\left(x - x_{prior}\right) \quad (8)$$

where C_{prior} expresses the uncertainty of the prior model state, a measure of our belief in the prior estimates, x_{prior}. The derivatives of this cost function are $J'_{prior}(x) = C_{prior}^{-1}(x - x_{prior})$; $J''_{prior}(x) = C_{prior}^{-1}$. A comparison of Eqs. (6) and (8) shows that this is really just another form of model constraint, with $M(x) = x_{prior}$, which can be achieved with the existing model constraint by setting $b = x_{prior}$. In practice, this allows us to enforce a prior belief in the distribution and range of the state vector elements (e.g. a climatology or physical or otherwise known 'reasonable' distributions), although only Gaussian distributions can be considered.

2.5. DA algorithm

The various constraints discussed above provide the cost function in Eq. (1) through their summation. This also applies to the derivatives $J'(x)$ and $J''(x)$. The cost function $J(x)$ is minimised using a gradient descent method (i.e. using $J'(x)$). Bounds are applied as a final constraint to the solution, to ensure that the state vector remains within physical limits. These can be supplied by the user. In the EO-LDAS prototype we use the limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) algorithm described in Byrd et al. (1995) and Zhu et al. (1997). In principle however, a number of different gradient descent algorithms could be used. The L-BFGS-B was selected for its efficient memory handling for high dimensional problems and the fact that it can optimise over a bounded domain, which is appropriate for this problem.

The algorithm then is quite straightforward: (i) read in configuration information and observations; (ii) provide an initial estimate of all state vector elements that we wish to estimate; (iii) iterate within the optimisation routine until convergence is reached (or using other criteria) to estimate the state vector; (iv) calculate the Hessian and then its inverse to provide the posterior covariance matrix, the estimate of uncertainty.

It is instructive to consider the contribution of these three terms in the estimates of Hessian matrix. The observational term can be ill-conditioned if the observations exhibit little sensitivity to some or all of the state variables, for example due to poor combinations of spectral and/or angular sampling. The addition of the prior and dynamic model terms then results in improved conditioning of $J''(x)$, as these extra terms compensate for the lack of observation sensitivity to some of the state variables. They also provide the ability to interpolate (i.e. rely more on the process model) between where we have observations. Importantly, the uncertainties are tracked throughout this process, so when e.g. interpolating over large gaps, we get the expected increase in uncertainty.

The DA system developed here can be viewed an extension of the methodologies that have been applied to inverting radiative transfer models by minimising a cost function. The addition of a linear dynamic model therefore only adds a handful of extra parameters to the problem (namely, the nature of the dynamic model itself and the associated covariance matrix, C_{model}, which may be simply diagonal). This is in a marked contrast with similar methodologies that either use a long time series of data for inverting one single parameter (in the case of inverting LAI as in Fang et al. (2008a, 2008b), Xiao et al. (2009) and Xiao et al. (2011). The temporal smoothness constraint is in itself an important feature, which is usually performed as a post-processing step after the parameter retrieval (Lu et al., 2007).

3. Experimentation

We present a series of experiments to demonstrate the operation of the EO-LDAS prototype and to explore the sorts of capabilities such a system could provide with data from the Sentinel-2 MSI sensor (see Table 1 for waveband information for Sentinel-2 MSI). The experiments use synthetic data for observations i.e. are derived from running the observation operator for a given state vector for what we suppose to be typical Sentinel-2 scenarios over one calendar year. We simulate top hat bandpass functions (1-nm sampling) according to the information in Table 1. The main aim of the experiments is to determine the improvement, in terms of reduced uncertainty, in biophysical parameter estimation that might be obtained by applying the EO-LDAS prototype for such scenarios. A subsidiary aim is to demonstrate the capability of the DA system to make predictions of data from a sensor not used in the DA process. Here, we do this by using the state vector estimates derived from the DA with synthetic MSI data, and make predictions of what a SPOT-5-like instrument would view (described below). These data are used in a cross-validation exercise within the experiments.

3.1. Experimental setup

In these experiments, we control the time trajectory of a subset of model parameters according to the functions given in Table 4, where t is the relative day of year (DOY) i.e. DOY normalised by 365. All other parameters take their default values given in Table 2. The functions for LAI and chlorophyll broadly mimic typical trajectories of these terms for crops: for LAI, a flat initial period, followed by a rise to maximum LAI and then a symmetric decrease; for Chlorophyll, a linear rise and decrease. The more arbitrary functions used for the soil brightness term s_1 we include to mimic rather broad variations over the year that might be supposed to be responses to soil moisture. A similar function is used here for leaf water, with a time lag of

Table 4
Upper and lower bounds for the state vector terms (in transformed space, where appropriate) used in the simulations, along with the temporal trajectory assumed.

#	Symbol	Lower limit	Upper limit	Temporal function
1	TLAI	0.067	0.995	LAI = 0.21 + 3.51 $\sin^5$(t)
4	TC_{ab}	0.135	1.0	C_{ab} = 10.5 + 208.7 t: t<=0.5 C_{ab} = 219.2 − 208.7 t: t>=0.5
6	TC_w	0.135	1.0	C_w = 0.068 + 0.020(sin(πt + 0.1) *sin(6πt + 0.1))
7	TC_{dm}	0.135	1.0	C_{dm} = 0.01
8	N	1	2.5	N = 1
9	s_1	0.001	0.4	s_1 = 0.20 + 0.18(sin(πt)*sin(6πt))

36.5 days. The quite large variation of these two latter terms is intended primarily to allow the operation of the data assimilation scheme to be explored over a wide range of conditions, rather than to too closely mimic some particular situation. In that context, the rather large time lag between soil brightness variation and leaf water content is unrealistic, but a larger phase between these terms should test the system to a greater extent than having all parameters following similar trajectories. Although the full set of state vector elements is 13 for each time sample, we attempt to retrieve only the 6 elements (numbers 1, 4, 6, 7, 8, 9 in Table 2) (per time sample) that we vary in these experiments, i.e. we assume the remaining elements fixed and known. This is partly to reduce the computational time required for the DA and more broadly because we believe it is sufficient to demonstrate the principles underlying the DA method. It is quite feasible to permit an estimation of 12 of the 13 elements (not the categorical variable directly through this method) but this is not the purpose of this exercise, and (arbitrary) variations in these additional terms would need to be defined to achieve this.

To approximate the Sentinel-2 MSI acquisition geometry (ESA, 2010), we assume one sample every 5 days (73 samples over the year), with a solar zenith angle corresponding to 10:30 local time at 50° N, random relative azimuth and random view zenith between 0° and 15°. Whilst these parameters do not provide a precise prediction of the likely MSI sampling and geometries, they are close enough to develop an understanding of the likely behaviour of the data. The random azimuth, for example, is clearly in error, but since the view zenith angle is so restricted, this will have very little impact; the local time at 50° N will in reality be slightly later than the nominal equatorial crossing time used here, but the details of the solar zenith angle are less important here than inducing a typical variation over the year (32° to 76° here). The simulation of one sample every 5 days mimics close to the maximum sampling achievable by MSI on 2 Sentinel platforms.

Synthetic observations were also generated for a SPOT-5 HRG-like instrument. This sensor has four wavebands (500–590, 610–680, 790–890 and 1530–1750 nm) (CEOS, 2010). We have assumed a revisit period of 13 days (to be out of sync with the synthetic MSI observations), although the differences are only up to two days from the MSI observations. The view zenith angle was limited to $+/-25°$ from nadir, with a random azimuth and a local overpass time of 10.30. In total, 28 observations were available in this dataset.

Uncorrelated Gaussian noise is added to the observations as part of the data synthesis. We assume the standard deviation of this to vary linearly from 0.008 at the shortest wavelength to 0.020 at the longest, for both the MSI and SPOT-5 HRG. These values are broadly twice those claimed for atmospheric correction of data from the NASA MODIS instrument (Roy et al., 2005). If an atmospheric 'correction' were performed on the data, we would generally expect the uncertainty in surface reflectance to be correlated across wavelengths, as e.g. an under-estimation of aerosol optical thickness would likely give rise to an over-estimate in reflectance for the shorter wavelength bands. Here, we have inflated the assumed (MODIS) uncertainties by a factor of two to take some account of such likely correlation. This highlights one of the benefits of ultimately using a more fully coupled surface-atmosphere observation operator, in that such features would fall naturally out of the model formulation and random noise might be more reasonably assumed for top of atmosphere radiance or reflectance. However, for the purposes of these experiments it is sufficient to treat only the surface (canopy-soil) elements of the observation operator.

We term this simulation set 'complete' for the purposes of this paper, in that it expresses a rather idealised situation where no clouds are present. A second synthetic observation set that we term 'cloudy' (36 observations for MSI and 15 for SPOT-5 HRG) is derived from this, for which we have removed 50% of the observations according to a correlated random function to mimic persistence of cloud cover. This induces ('cloud') data gaps of up to 60 days (mean gap 10.3 days, standard deviation 12.6 days for MSI).

As noted above, the cost function minimisation is achieved in EO-LDAS with the L-BFGS-B algorithm. A bounded minimisation is performed within this code, with the limits specified on the (transformed) state variables given in Table 2 (transformations in Table 3). Thus, all state variable estimations below proceed with the prior knowledge of an upper and lower bound. There are several convergence criteria that can be used with the L-BFGS-B, including an absolute threshold on the cost function and a relative (per iteration) threshold. In all experiments, these are set to low values, which means that more iterations might be employed than strictly necessary in any operational context, but making sure that the global minimum (or very close to it) is reached in each estimation. Because of the additional costs of processing full bandpass functions, all 'initial' processing is performed using the median (1 nm) wavelength of each waveband. A 'polishing' step is then performed to achieve convergence from this starting value, using the full bandpass sampling. The effect of applying the full bandpass functions tends to be generally quite minor.

We have initially tested the system without observational noise and confirm that the scheme retrieves the truth to within the bounds implied by the convergence criteria and machine precision. Processing time for a single set of 73 time samples with MSI spectral sampling, solving for 6 state vector elements for each day of the year (2190 in total), is currently several hours on a 3 GHz Intel processor on a single core, but this is partially due to very stringent convergence criteria used whilst testing the code and partially because this prototype implementation requires some significant efforts in computer code optimisation.

In all experiments, we set the prior estimate of the state vector to the values shown in Table 2, with very large diagonal uncertainty terms. This effectively removes the prior constraint from consideration in these experiments, as we wish to conduct experiments based only on model and observational constraints here.

In the following sections, we examine the result of applying the weak constraint variational data assimilation approach described above to the synthetic dataset. For all cases, we assume that the uncertainty in the observations is known and that it is Gaussian and uncorrelated between wavebands and between dates. In the first case (3.2), we solve for state vector estimates assuming no dynamic model constraint other than the weak prior (standard deviation 8). This acts as a baseline for further experiments. In the second case (3.3) we assume that model uncertainty γ is unknown and attempt to solve for it and the state vector for each day in the year with a form of cross-validation exercise using the SPOT-5 HRV synthetic observations. The 'true' values of γ for individual state vector elements are shown in Table 5. The DA is performed with the 'complete' (i.e. 5-day sampling MSI) dataset in that case. Finally, we repeat that experiment for the 'cloudy' dataset (3.4).

Graphical results (Figs. 1–2) are presented as untransformed biophysical variables (i.e. LAI, C_{ab}, C_w, C_{dm} , N and s_1), showing: the 'true' ('original') state vector (dashed line); circles and error bars (shaded region) shows mean and 95% credible interval bounds (at plus/minus 1.96 standard deviations). We will term 1.96 standard deviations 'uncertainty' for the remainder of the paper, unless the statement is otherwise qualified. The uncertainty bounds are slightly

Table 5
Model uncertainty γ for each parameter, calculated from the synthetic model state vector. TC_{dm} and N were kept constant, so there is no theoretical model uncertainty associated.

#	Symbol	First difference uncertainty	Second difference uncertainty
1	TLAI	188	8298
4	TC_{ab}	303	7315
6	TC_w	132	2277
9	s_1	212	3861

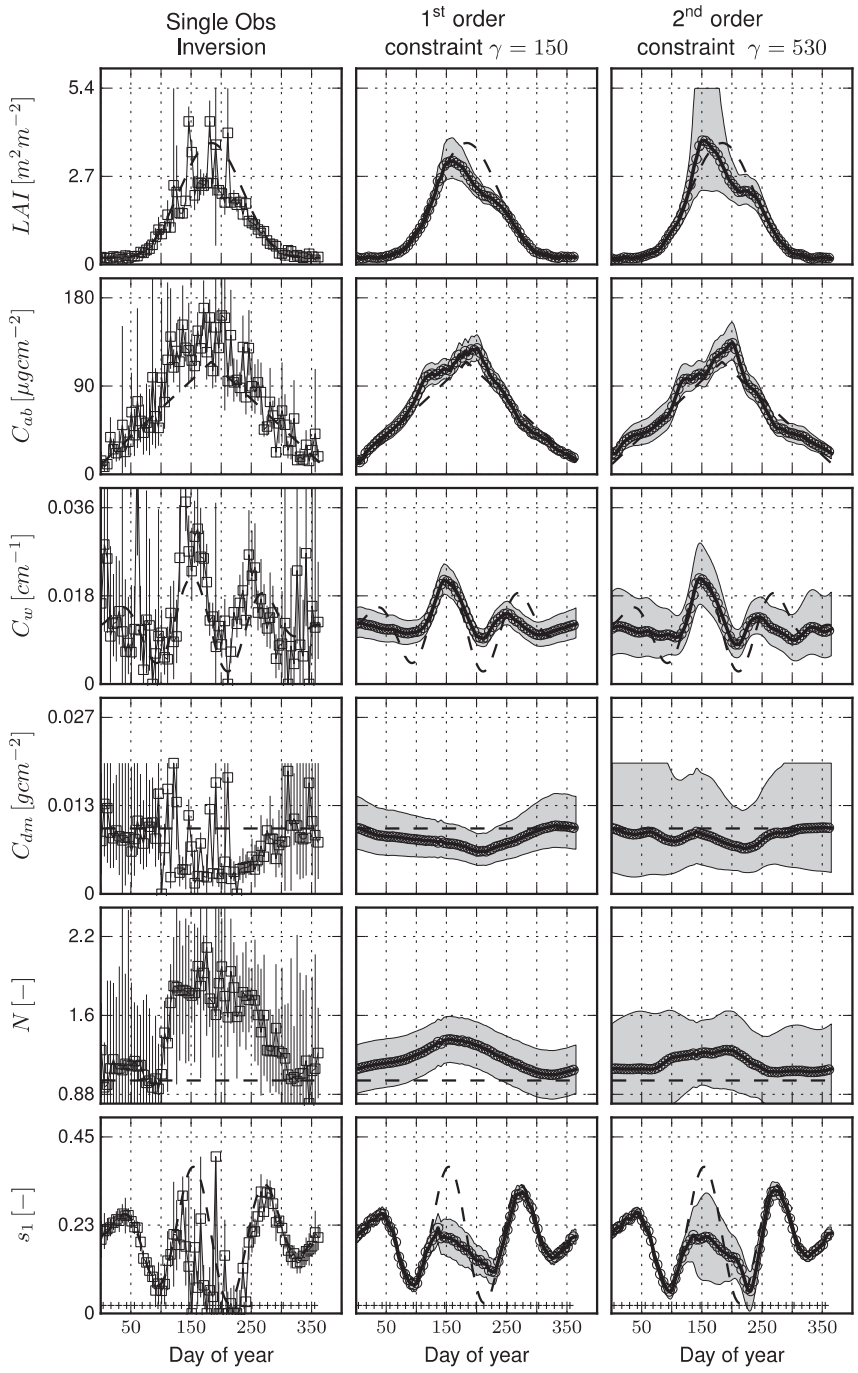

Fig. 1. Base level state vector estimated from inverting single observations, (left column) and for model uncertainty unknown and estimated through cross validation—first difference constraint (central column) and second difference constraint (third column). Results for each of the six parameters are shown in rows. True values are shown as a dashed line. The full lines are the posterior means, and the shaded area represents the associated +/−1.96 standard deviations interval. MSI observations are shown as open symbols. Crosses along the bottom row indicate the location of the cross validation acquisition dates.

larger for the upper limits than for the lower limits (other than for N and s_1) due to the nature of the transformations used in the approximate linearization (Table 3). Tabular results for the experiments (Tables 6–9) are expressed in transformed parameter space, as that is the space in which the state vector is inferred and in which the Gaussian statistics derived are most natural.

3.2. Baseline estimates

We first produce a baseline estimate of the six state variables over the 73 time periods in the year, assuming no constraint to the solution other than the bounds noted above, the (noisy) observations, knowledge of the uncertainty in the observations, and the weak prior constraint.

The results for the baseline experiment are produced using the EO-LDAS system with each observation set (i.e. all wavebands, but only one angular sample) independently. The algorithm requires an initial guess of the state vector and iterates to its final estimate. The initial estimate of the state vector in each case and all subsequent estimates is taken as the value used in the prior constraint.

In Fig. 1, the column titled 'single obs inversion' shows the results of this state vector estimate for the six parameters that are varied,

Table 6
Mean posterior uncertainty. Figures refer to the complete daily time series, while figures in brackets refer to the mean posterior uncertainty only considering the dates where observations are available.

Symbol	Non-cloudy			Cloudy		
	Uncertainty Single obs.	Uncertainty 1st diff	Uncertainty 2nd diff	Uncertainty Single obs.	Uncertainty 1st diff	Uncertainty 2nd diff
TLAI	0.18 (0.05)	0.04 (0.04)	0.06 (0.06)	0.21 (0.05)	0.06 (0.05)	0.09 (0.07)
TC_{ab}	0.20 (0.10)	0.04 (0.04)	0.06 (0.06)	0.22 (0.09)	0.06 (0.05)	0.08 (0.06)
TC_w	0.23 (0.18)	0.07 (0.07)	0.13 (0.13)	0.24 (0.19)	0.10 (0.10)	0.17 (0.16)
TC_{dm}	0.24 (0.22)	0.13 (0.13)	0.28 (0.28)	0.24 (0.23)	0.19 (0.19)	0.36 (0.35)
N	0.29 (0.55)	0.21 (0.21)	0.37 (0.37)	0.27 (0.55)	0.32 (0.32)	0.44 (0.40)
s_1	0.17 (0.04)	0.02 (0.02)	0.03 (0.03)	0.20 (0.04)	0.04 (0.03)	0.05 (0.03)

transformed back to their biophysical meanings (through the inverse of the functions in Table 3). The sub-plots rows show results for the observation operator parameters LAI, C_{ab}, C_w C_{dm}, N and s_1 respectively. The uncertainty (average credible interval) associated with each (transformed) parameter for the baseline experiment is given in Table 6 ('single obs.'). Relating these uncertainties to the parameter ranges (Table 4), we note that they are around 5% for TLAI, 10% for s_1 and TC_{ab} respectively for dates where there are observations, more than 20% for (transformed) leaf water and dry matter content and around 33% for N. We can suppose these then to be typical uncertainty values for MSI sampling (with the assumed noise characteristics). The cross correlations associated with these, illustrated in Table 7 are highly variable from one sample to the next. The median values given show quite strong negative correlations between TLAI and TC_{ab} and TC_w but positive correlations with s_1 and TC_{dm}. The median s_1 shows negative correlations with all terms other than TLAI. Despite the fact that the average transformed LAI uncertainty is only around 5%, we can see if Fig. 1 that both the error and uncertainty can be rather high. Around peak LAI, results from individual samples vary by around LAI 2.5 and there is a general tendency to underestimate. The general trends of C_{ab} and C_w are discernable, but there is large variation and large uncertainty. The terms that are supposed to be constant here, C_{dm} and N depart significantly from their true state and the negative correlation is evident in the state trajectories around the central part of the year.

How then can we improve on this situation? The ways to reduce uncertainty are to have data with lower noise characteristics, to average or smooth in some way, or to add other constraints to the solution. In any realistic scenario, we have only limited control of the first of these. Averaging and smoothing then are the general pragmatic responses to such issues. If however this is performed *ad hoc* as a post-processing step to any individual term (e.g. only LAI) this would not take account of the cross correlation in the uncertainties which can only give sub-optimal results.

In spite of these quite high levels of uncertainty (and correlation of uncertainty) for these estimates, there is clearly quite a strong correlation between the values of the state vector and its neighbours in time. The general underlying patterns are apparent in the 'complete' scenario, although much of the (potentially important) detail will be lost in a more realistic 'cloudy' scenario. The enhancement of this temporal correlation effect and the suppression of the noise are at the heart of all regularisation approaches and the essence of weak constraint data assimilation. If we have some model ('expectation') of the temporal trajectory of the state vector, then we can use this to filter the unwanted noise. As noted above, this may be a model based on our understanding of radiation interception and biogeochemical cycling (e.g. Quaife et al., 2008) driven by some set of external (environmental) parameters, or it may simply be some parametric curve that we believe can mimic e.g. the phenological development of LAI. In either case, what DA aims to achieve is an optimal merging of such models (through the adjustment of the state vector or essentially a calibration of the parameters controlling the development of state in the model) and the observations. For land surface monitoring, there are several options for such models for LAI development as mentioned, and up to a point for some other state variables (e.g. soil moisture), but there is very little to guide information extraction on many other state variables that affect the observations (e.g. leaf chlorophyll concentration or dry matter). In such a case, we need to develop simple methods, within a DA framework. Fortunately, there are many to choose from, although as Twomey (2002) points out, the results are likely to be similar for most of these methods: indeed, it would be worrying if they were not.

Table 7
Single observation posterior correlation matrix. Elements above the main diagonal show the results for DoY 186, whereas the elements below the main diagonal represent the median of all dates.

Symbol	TLAI	TC_{ab}	TC_w	TC_{dm}	N	s_1
TLAI	1.00	0.16	−0.05	0.47	−0.25	0.58
TC_{ab}	−0.44	1.00	0.15	−0.11	−0.47	0.34
TC_w	−0.42	0.35	1.00	0.04	0.01	−0.14
TC_{dm}	0.30	0.27	−0.27	1.00	0.42	−0.36
N	0.00	−0.21	0.07	−0.43	1.00	−0.85
s_1	0.76	−0.53	−0.40	−0.25	−0.28	1.00

3.3. DA: Complete scenario

Here, we apply first order and second-order derivative constraints to the solution, but we expect the results to be broadly similar. In both cases, we need only supply some estimate of the uncertainty associated with these constraints through the smoothness term γ to achieve a regularised solution to the state vector estimate. These constraints are applied by incorporating a model that, in the absence of any observations, would set the first (second) derivatives of the state variables to zero. Assuming that we apply the same (strength of) constraint to the whole time series, we need to supply an estimate of the mean squared first (second) difference in the parameter values (true values in Table 5). For the first (second) difference then, this can be thought of as an estimate of the uncertainty in a zero-order (first order) process model over one time step as noted above.

We use a form of cross validation to estimate γ. This is achieved with a synthetic dataset from an alternative SPOT-5 HRG-like sensor. The core of the exercise then is a comparison between these synthetic data (driven by the 'true' values of the state vector, plus random noise as above) and a simulation of the same sensor wavebands and acquisition geometry driven by the state vector estimated from the synthetic data from Sentinel-2 MSI. We choose this cross validation sensor as one different to MSI to stress that one role of a DA system of this sort can be to provide simulated data of sensors other than those used in the DA exercise. Here, we measure the average squared difference between the synthetic HRG data and the DA simulated observations, weighted by the uncertainty in the synthetic data, and term this RMSE in cross validation. The locations of the synthetic HRG observations are indicated in the lower panel of Fig. 1 by + symbols.

Fig. 3 shows the error in cross validation as a function of γ for the model first- and second-order difference constraints for the complete case (black circles and squares respectively). There are clear minima for these functions, which provide estimates of the optimal model uncertainty (averaged over all terms). Also shown in the figure is a set of vertical lines that represent the theoretical value of the smoothness term for each of the state vector elements that vary over time (from Table 5). For the first order constraint, the minimum of the cross validation function is $\gamma = 150$ which is very close to the theoretical values. For the second-order constraint the cross validation RMSE minimum at $\gamma = 530$ is rather less than the theoretical values. For both cases however, we observe a very broad minimum, so there is quite a large range of values of γ that allow almost equally good prediction of the synthetic cross validation HRG observations.

Table 8 provides statistics on the uncertainty reduction, (the posterior uncertainty estimate from the DA relative to that after solving for each sample separately and assuming the prior uncertainty where there are no observations). The average improvement in uncertainty is 4.07 for the first order constraint and 2.73 for the second-order difference constraint. This is very significant but it must be remembered that 4/5 of the samples in the 'single obs' solution have only the prior constraint and uncertainty. Examining only locations where observation lie (i.e. ignoring interpolation performance relative to the *a priori* estimate), we see the uncertainty reduction drop by nearly 50% in this case, down to 2.20 for the first order constraint and 1.30 for the second difference constraint. From those figures, we would suppose the first order constraint to be greatly superior to the second-order constraint, but if we look at the plots in Fig. 1, the second-order constraint results seem to have more reasonable uncertainty bounds than the other results. This is at least partially because the apparent uncertainty resulting from the DA is strongly dependent on the value of γ used in the model constraint: the higher the value of γ, the smoother will be the solution and the lower the estimate of uncertainty. The only check we have done on the *veracity* of the solution comes from the cross validation, which is an indirect check: in any non-synthetic experiment we rarely know the 'truth' to any great degree of certainty. Since we have a synthetic experiment here, we can however test how frequently the derived solution matches the (synthetic) truth within the claimed uncertainty bounds. One reasonable summary measure of this is the percentage of true values of state vector elements that lie within the 95% credible interval claimed by the DA results. These are shown in Table 8. We can see that for the 'single obs' estimates (no regularisation), only around 64% of the state vector lies within the 95% credible interval claimed by the solution. The figure is as low as 58% for TC_{dm}. We can suppose the average estimated uncertainty then to be only around 67% of the true value, i.e. we should inflate the estimated uncertainty by a factor of around 1.5. This would apply equally to the results in Table 6. We see almost the same value for the first order constraint, which suggests the reduction in uncertainty by a factor of 2.2 is likely true. For the second-order difference constraint however, around 84% (Table 8) of the sample lie within the uncertainty bounds, so here, a better estimate of the uncertainty reduction might be around 1.70 rather than the 1.30 reported. This apparent under-reporting of the uncertainty is worthy of comment and there could be several reasons for this. One explanation could be that we are simply under-estimating the uncertainty from the approximations made when calculating the Hessian for the observation operator. A more likely reason is non-linear effects in the treatment of uncertainties. In spite of our attempt to account for gross non-linear impacts through parameter transformations, residual non-linear effects may be causing this under-estimation of uncertainty by a factor of around 1.5.

3.4. DA: Cloudy scenario

Fig. 2 shows the DA results for the cloudy scenario. This is a much more realistic test for a DA system. The task now is not only to reduce the uncertainty at the points where we have observations but also to try to provide an effective interpolation over data gaps. The cross validation plots for this case are shown in Fig. 3 (white circle and square) and provide a much more narrow minimum. This implies that to achieve acceptable results in cross validation, the range of γ values that can be tolerated is much more restricted. The minima of these functions however are well within the bounds of the cross validation results for the 'complete' scenario and the optimal γ indicated very similar to that obtained from the previous results. This indicates that the method for estimating γ is quite robust, even when there are large data gaps. Unsurprisingly, the absolute value of the cross validation RMSE is higher for the cloudy case, indicating poorer performance in prediction for this lower quality dataset.

Table 9 shows the reduction in uncertainty for this experiment. One striking feature of these is that the percentage of cases within the credible interval is now above 80% in both cases, meaning that the reported uncertainties are close to the true values. Whilst the apparent reduction in uncertainty is apparently quite small (indeed, there is an increase in uncertainty for some state vector elements) at 1.53 for the first order constraint and 1.14 for the second order, when weighed against the improved statistical representation, these rise to values directly comparable with the results from the previous experiment. The credible intervals shown in Fig. 3 are now realistic representations of the state vector elements and their uncertainties, achieved with only 50% of the samples of the previous experiment and with large data gaps, which is an important result.

Fig. 4 shows the posterior correlation matrices (the inverse Hessian matrix) for the cloudy scenario. The general pattern of this matrix for the 'complete' scenario is rather similar so not shown here. Obviously, the correlation is unity along the leading diagonal. Another important feature is that the broad patterns of positive and negative correlations that we noted for the 'single obs' solutions remains here. There is negative correlation between s_1 and all terms by TLAI. There is negative correlation between TLAI and TC_{ab} and TC_w but

Table 8
Uncertainty reduction relative to the single observation inversion, as well as percentage of cases where the true parameter lies within the estimated 95% confidence interval. Results for non-cloudy scenario, complete time series.

		Complete time series				Observations only				
#	Symbol	Unc. red 1st diff	Unc. red. 2nd diff	% cases (1st diff)	% cases (2nd diff)	Unc. red 1st diff	Unc. red. 2nd diff	% cases (1st diff)	% cases (2nd diff)	% cases (single)
1	TLAI	4.89	2.96	75.3	90.4	1.44	0.85	72.6	91.8	63.0
4	TC_{ab}	5.24	3.58	61.1	65.2	2.57	1.74	60.3	65.8	65.8
6	TC_w	3.47	1.77	51.2	69.9	2.72	1.38	50.7	71.2	60.3
7	TC_{dm}	1.82	0.85	87.7	100.0	1.64	0.76	87.7	100.0	57.5
8	N	1.40	0.79	59.2	100.0	2.67	1.50	58.9	100.0	60.3
9	s_1	7.59	6.43	67.1	72.3	2.13	1.56	63.0	72.6	75.3
	Mean	4.07	2.73	66.9	83.0	2.20	1.30	65.5	83.6	63.7

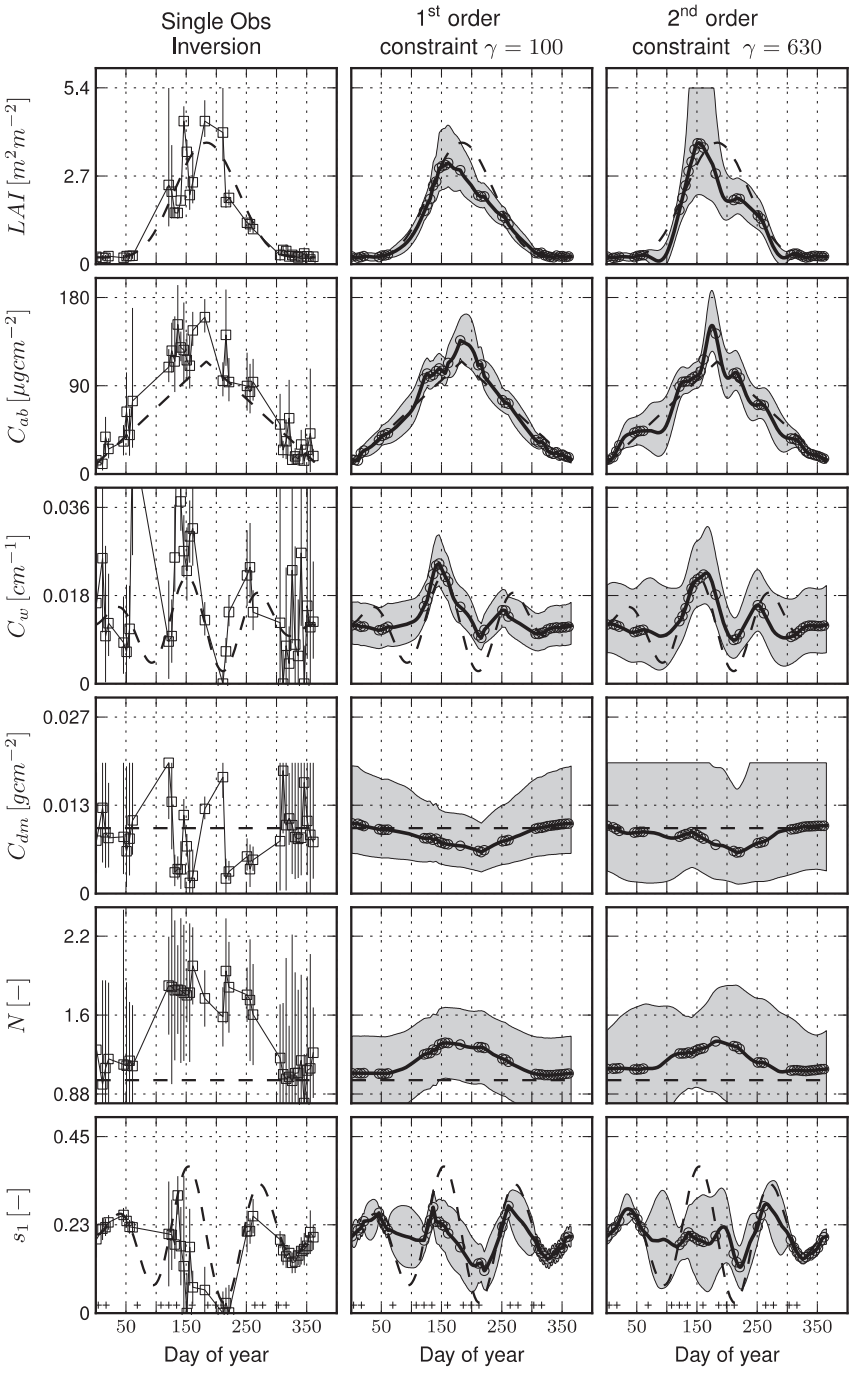

Fig. 2. Base level state vector estimated from inverting single observations, (left column) and for model uncertainty unknown and estimated through cross validation—first difference constraint (central column) and second difference constraint (third column). Reduced number of acquisitions due to cloud cover scenario. Results for each of the six parameters are shown in rows. True values are shown as a dashed line. The full lines are the posterior means, and the shaded area represents the associated +/−1.96 standard deviations interval. MSI observations are shown as open symbols. Crosses along the bottom row indicate the location of the cross validation acquisition dates.

positive correlation with TC_{dm}. These patterns are consistent for both constraints used. We notice then that the application of the dynamic model (regularisation) in time does not remove the correlations arising from the inverse Hessian of the observation cost function, but rather it 'spreads' uncertainty correlation out in the time domain. This is particularly visible in the second-order constraint matrix in Fig. 4 where we can clearly see this smoothing being greater where there are data gaps (s_1 is a good example of that). Equally, where a part of the state vector has been strongly influenced by the regularisation (e.g. N for the first order constraint) we see very high correlation at all time steps. Another interesting feature of this figure is the fact that for some state vector elements (e.g. N for the second-order constraint) we can clearly see the influence of the periodic boundary condition).

4. Discussion

4.1. The value of an EO-LDAS

This paper outlines a scheme for a weak constraint data assimilation system, developed in the ESA EO-LDAS project, designed for integrating Earth Observation data from a variety of sources over

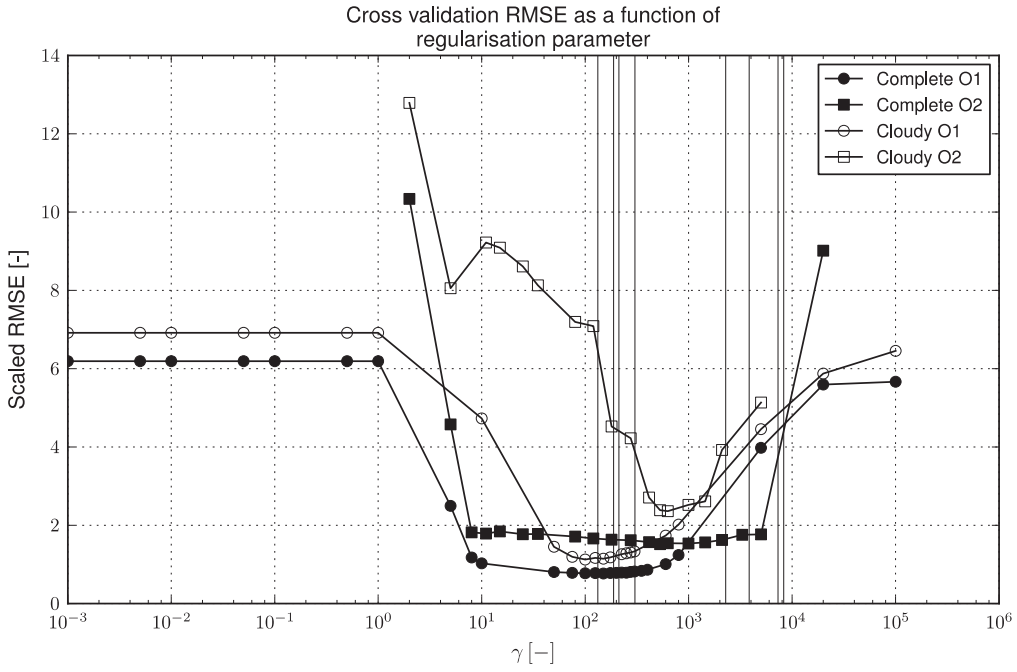

Fig. 3. Error in cross validation scaled by observational uncertainty for varying model uncertainty γ for first and second-order constraints. Vertical lines around 200 represent the theoretical value of γ for each of the 4 time-varying state variables using a first order constraint, and vertical lines around 5000 represent the theoretical values for γ for each of the 4 time-varying state variables using second-order constraint.

arbitrary time scales, and through that to multiple spatial resolutions. It has the direct potential to be extended to spatial constraints, although this is not explored here. The scheme is designed to allow interface with process models, should they be available, though only an empirical regularisation model is shown in this paper. The core of the system is a set of constraints on: (i) prior estimates of the state vector; (ii) a linear model of the state vector; (iii) observation operator (RT model) predictions of a set of EO data and a DA scheme around these using an iterative bounded optimisation approach (L-BFGS-B).

In this paper, we have set up and run a synthetic data experiment with EO data mimicking those that might be provided by the MSI sensors on the forthcoming Sentinel-2 platforms. Experiments in DA are conducted for an idealised 'full coverage' scenario (5-day sampling) and for a 'cloudy' case (average around 10-day sampling but with large data gaps of up to 60 days). The results are compared to baseline experiments where we attempt to estimate the state variable trajectories over the course of a year for a subset of the total state variables (six elements per observation period). The prior term is used only very weakly here, although bounds are set to the state vector elements. Further, we assume that we have direct access to the surface reflectance (as opposed to top of atmosphere radiance), and that the noise on the observations is uncorrelated and of known magnitude. Broadly however, we can claim that the baseline results should be indicative of those that might be obtained from Sentinel-2 data using 'traditional' estimation methods. For what we suppose to be a typical observation noise scenario, the uncertainty can be a quite large proportion of the signal for important terms such as LAI, this for a peak LAI of only around 3.7, although on average the uncertainty in TLAI may only be around 5%. This then, relates to the information content of a single MSI observation for this level of noise, assuming some important terms such as leaf angle distribution are known precisely. These results are not surprising but are simply a manifestation of the difficulty of the inference of biophysical parameters from radiometric observations: the problem may often be ill-posed (consider the situation if only two wavebands at red and near infrared were available), but even if it is not strictly so, there may not be sufficient information to very well constrain the information we require. In any case, there can be quite high correlation in uncertainty.

The ways to improve this situation are: (i) to obtain more observations (although more observations do not always translate in more information: consider again sampling at only red and near infrared wavelengths in trying to constrain e.g. leaf water content); (ii) to add some other forms of information; or (iii) average the data. Much *a priori* information has been used in the past to help constrain these problems, but this has often been approached in a rather *ad hoc* manner. Examples include: assuming some terms known,

Table 9
Uncertainty reduction relative to the single observation, as well as percentage of cases where the true parameter lies within the estimated 95% confidence interval. Results for cloudy scenario.

		Complete time series				Observations only				
#	Symbol	Unc. red 1st diff	Unc. red. 2nd diff	% cases (1st diff)	% cases (2nd diff)	Unc. red 1st diff	Unc. red. 2nd diff	% cases (1st diff)	% cases (2nd diff)	% cases (single)
1	TLAI	3.33	2.39	82.7	74	0.965	0.68	88.9	91.7	63.9
4	TC_{ab}	3.68	2.93	80.0	89.6	1.82	1.58	61.1	83.3	58.3
6	TC_{w}	2.33	1.4	64.1	85.2	1.90	1.18	83.3	88.9	61.1
7	TC_{dm}	1.25	0.669	100	100	1.18	0.656	100	100	58.3
8	N	0.835	0.597	91	100	1.73	1.38	88.9	100	58.3
9	s_1	4.65	4.53	63.6	78.1	1.57	1.35	69.4	72.2	75.0
	Mean	2.68	2.09	80.2	87.8	1.53	1.14	82.0	89.3	62.5

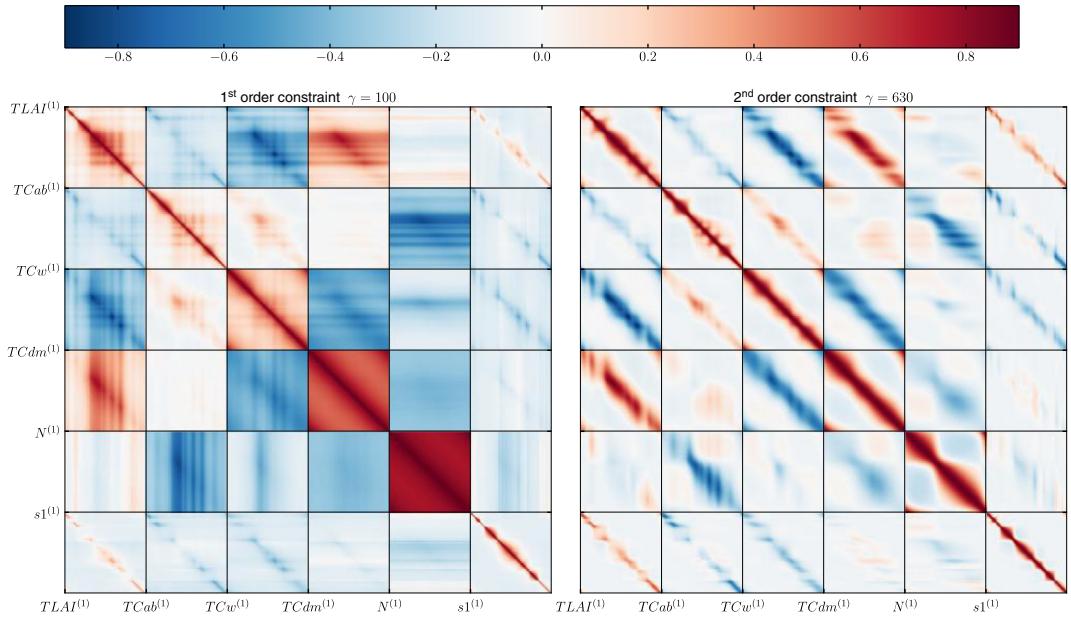

Fig. 4. Posterior correlation matrices for the cloudy scenario. Labels indicate the location of the first day for component of the state vector.

without considering the impact of uncertainty in these, or imposing degrees of smoothness; assuming that some terms are constant over some arbitrary time period; or *post hoc* low pass filtering to the final results. Given its success in other field of science and engineering, many authors have proposed that DA should be seen as the route to integration of the various forms of information one might wish to use to constrain the estimation. Key to DA is the weighting between the various sources of evidence, and key to this is assigning uncertainty correctly to the sources. This is a feature of the approach that dramatically differentiates it from the way in which VIs are mostly used in EO. As we note in the introduction, if we wish to estimate biophysical parameters (such as LAI) there is generally some form of calibration (against ground observations or RT model runs) but it is extremely rare that those model uncertainties are considered in mapping the product. Other processing steps such as angular normalisation may have taken place, but again, any concepts of uncertainty arising from these are on the whole disregarded. All of these issues *could* be addressed within a DA framework, even if the source of the EO information were to be VIs.

If a biophysical process model is available to predict the development of the state variables that control the remote sensing signal, this can clearly add information to help constrain the problem. If information from the observations feed back to improve the estimates of the parameters controlling the process model or alternatively improve the state estimates, then a better integration of observations and model is achieved, which will likely better constrain additional terms estimated by the process model. This has been argued by Quaife et al. (2008) and others who have worked on integrating EO data and e.g. Carbon flux process models. However, models such as these simply do not provide information on a large number of the variables that affect EO signals, and this is likely to remain the case for the foreseeable future. Exercises in EO-model integration then understandably tend to focus of the points of common linkage (which often is no more than LAI, being supposed linearly related to leaf Carbon) and then applying the 'traditional' methods to the remaining parameters (assuming them known or at best constant over time). In this paper, and in the EO-LDAS work in general, we have taken the focus away from working with some specific process model, and tried to consider the more general case and the sorts of constraints that might be appropriate. If no physical model is available, empirical concepts of smoothness in the state variables come to the fore. These ideas become even more important if one considers constraint in the spatial domain, where physical or even biological process models are almost completely lacking to aid biophysical parameter estimation.

The EO-LDAS scheme that we have built is capable of using any linearised process model and of more general interface to process model codes provided the cost function and its derivatives can be calculated. In the prototype and in this paper we have examined first- and second-order derivative constraints as general, appropriate (empirical) models for biophysical parameter estimation in DA. We have simulated typical profiles of LAI and leaf chlorophyll concentration and rather complex profiles of leaf water concentration and soil brightness and shown that with Sentinel-2 MSI data every 5 days, a reduction in uncertainty by a factor of around 2 might generally be achieved. More interestingly perhaps, after compensation for errors in uncertainty prediction, we saw that similar reductions might be achieved even when there are large data gaps and 50% of the samples lost due to cloud cover.

We have also demonstrated (Fig. 2) that it is feasible to estimate the required hyperparameters from some form of cross-validation exercise to impose an appropriate degree of model uncertainty, and that quite consistent results can be obtained even under cloudy conditions. This is an important practical point for the eventual operationalisation of these methods, but the area requires a little more discussion of practical issues in its implementation.

Approximate linearization of the RT model variables here, following Weiss et al. (2000), has allowed Gaussian distributions to be assumed throughout. Although we have not directly investigated any residual non-linear effects in this study, some evidence is provided that on average we may be predicting only around 2/3 of the true uncertainty.

4.2. Future directions

In this paper, we have only demonstrated DA for a homogeneous observation system, i.e. one for which we have assumed the spectral sampling (and in effect, spatial resolution) for all observations is the same. Using the EO-LDAS prototype for spectrally heterogeneous systems is straightforward, but further work is needed to test the multi-scale concepts that would more generally be required. Within the existing prototype, the state vector can represent any mixture of temporal or spatial samples. The concepts of temporal smoothness used

here apply equally to the spatial domain (indeed, such ideas form the basis of the field of geostatistics (e.g. Atkinson and Lewis, 2000)), so the prototype can be used directly to link a state vector representation on a spatial grid, via appropriate specification of the matrix *A*. Indeed, one could consider the experiments performed in this paper simply as being on a spatial transect, rather than as we have assumed a temporal sampling pattern. The only practical difference is that in that case, the viewing and illumination angles would be near identical for all samples.

The EO-LDAS prototype is designed to allow a (relatively) large number of state variables to be estimated simultaneously in a variational system (>2000 demonstrated here). One potential advantage of this is that information can be passed between any of the state vector elements. In practice, we have only used rather local information transfer in the model constraints applied here (differences with neighbours in time) and this approach could also be implemented as a sequential smoother. In viewing the temporal experiment we have performed as effectively equivalent to a spatial experiment, the neighbourhood need not be very different (i.e., in the spatial sense, we could follow the approach here and directly connect information in one grid cell to its 8 neighbours). However, this variational system maintains the capacity for more distant (time or space) connections, for example in applying multiple scale constraint.

A point that we have not dwelt on in this paper is the time required for processing. This is currently around several hours for solving for >2000 state vector elements using 73 samples for what equates to a single pixel (albeit for all samples over a year). The experiments in this paper were conducted over around 120 UNIX cores, so quite large-scale experiments are feasible using University computing resources. Clearly the processing requirements would need to be greatly reduced if such a system were to be proposed for operational processing. The computer code is not on the whole written to be fast, but rather to be adequate to learn about using this form of DA. There are various ways in which this might be tackled: clearly the very tight convergence criteria could be somewhat relaxed, and more efficient codes could be written, but there will always be a relatively large overhead on multiple calculations of a radiative transfer model. Pragmatic ways to overcome this issue have mainly in the past dealt with using LUTs or ANNs to sample or approximate the observation operator, but clearly in the DA framework we must consider representational error in any such emulation. One avenue that holds much promise is that of Gaussian Process (GP) emulators (Kennedy & O'Hagan, 2000; 2001), a form of regression that has been successfully used to simulate computationally costly models runs through simple functional approximations. The great benefit of this latter approach is that uncertainties in the emulated model are included and that derivatives of the model can also be easily produced. If we consider the observation operator as a sampled function with GP emulation, it is interesting to note that the underlying concepts implying smooth interpolation with treatment of representation uncertainty are of course the same as we are performing in the temporal (or indeed spatial) process model in the DA.

5. Conclusions

The EO-LDAS prototype that is described in this paper has been demonstrated to be capable of simultaneously estimating a state vector of over 2000 elements of surface biophysical characteristics in a synthetic experiment using simulated Sentinel-2 MSI data. Although the processing time required for this is currently substantial, this is a significant step in the size of such problems that can be tackled simultaneously. The ability to do this derives from the use of AD-generated adjoint code for the observation operator at the heart of the DA system.

The DA scheme that has been developed is a weak constraint variational system. The value of such a scheme has been demonstrated using the synthetic MSI data to show a reduction in uncertainty of up to around 2 when a linear dynamic model is used in the DA. The linear dynamic model is proposed as a general implementation that can potentially be interfaced to biophysical process models through linearization. It is used in this paper with first and second-order derivative constraints (zero- and first-order process models) which are shown to be sufficient to track rather complex biophysical parameter trajectories via a radiative transfer model 'observation operator' interface to the synthetic EO data.

We have noted at various points in this text, that some aspects of the EO-LDAS prototype are still under development of testing, but what actually is provided by the prototype code is a functioning tool for exploring many issues in DA and for estimating information on surface biophysical parameters. The tool is designed as a weak constraint variational system, but we have argued that it can also be used sequentially as it stands. We have demonstrated the use of the tool and of DA concepts in reducing uncertainty in biophysical parameter estimation in a temporal sense, but also argued the equivalence of this (in DA in general, but in the tool specifically) for the spatial domain as well. We have used only empirical 'regularisation' concepts in demonstrating the DA, but noted that these are powerful general concepts that are extremely useful, particularly if biophysical models do not treat some of the parameters we are concerned with. In the more general case though, any linearization of a more process-driven model can be directly interfaced to the EO-LDAS prototype.

There is clearly quite a long way to go from initial experiments with relatively slow computer codes to an operational system for land data information extraction from EO, i.e. an operational EO-LDAS, but the concepts explored here demonstrate the power and potential flexibility of such an approach.

Acknowledgements

We gratefully acknowledge the support of ESA through the EO-LDAS project 22205/09/I-EC for funding this work. We also acknowledge the support of the (UK) National Environment Research Council (NERC) National Centre for Earth Observation (NCEO) for its support of several of the personnel involved in this work. We would further like to thank the attendees of the EO-LDAS Community Workshop held in ESA ESRIN in November 2009 for their feedback and inputs to this study. We also thank the anonymous reviewers for their helpful comments.

References

Aschbacher, J., & Pérez, M. P. M. (2010). GMES—Status review and policy developments. In K. Schrogl, W. Rathgeber, B. Baranes, & C. Venet (Eds.), *Yearbook on Space Policy 2008/2009* (pp. 188–207). Vienna: Springer Vienna Retrieved from. http://www.springerlink.com/content/k546p12832126066/

Asrar, G., Kanemasu, E. T., Jackson, R. D., & Pinter, P. J., Jr. (1985). Estimation of total above-ground phytomass production using remotely sensed data. *Remote Sensing of Environment*, *17*(3), 211–220.

Atkinson, P. M., & Lewis, P. (2000). Geostatistical classification for remote sensing: An introduction. *Computers & Geosciences*, *26*(4), 361–371, doi:10.1016/S0098-3004(99)00117-X.

Baret, F., & Guyot, G. (1991). Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment*, *35*(2–3), 161–173.

Behrenfeld, M. J., Randerson, J. T., McClain, C. R., Feldman, G. C., Los, S. O., Tucker, C. J., et al. (2001). Biospheric primary production during an ENSO transition. *Science*, *291*(5513), 2594.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

CEOS (2010). The CEOS constellation for land surface imaging. Retrieved December 15, 2010, from. http://wgiss.ceos.org/lsip/lsic.shtml

Chen, J. M., Pavlic, G., Brown, L., Cihlar, J., Leblanc, S. G., White, H. P., et al. (2002). Derivation and validation of Canada-wide coarse-resolution leaf area index maps using high-resolution satellite imagery and ground measurements. *Remote Sensing of Environment*, *80*(1), 165–184.

Choudhury, B. J. (1987). Relationships between vegetation indices, radiation absorption, and net photosynthesis evaluated by a sensitivity analysis. *Remote Sensing of Environment*, *22*(2), 209–233.

Clerici, M., Vossbeck, M., Pinty, B., Kaminski, T., Taberner, M., Lavergne, T., et al. (2010). Consolidating the two-stream inversion package (JRC-TIP) to retrieve land surface parameters from albedo products. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, *3*(3), 286–295.

Combal, B., Baret, F., Weiss, M., Trubuil, A., Mace, D., Pragnere, A., et al. (2003). Retrieval of canopy biophysical variables from bidirectional reflectance: Using prior information to solve the ill-posed inverse problem. *Remote Sensing of Environment*, *84*(1), 1–15.

Council of the European Union (2010). *Taking forward the European Space Policy*. .

DMCII (2010). DMC constellation. Retrieved December 15, 2010, from. http://www.dmcii.com/about_us_constellation.htm

Eilers, P. H. C. (2003). A Perfect Smoother. *Analytical Chemistry*, *75*(14), 3631–3636, doi:10.1021/ac034173t.

Enting, I. G. (2002). *Inverse problems in atmospheric constituent transport.* : Cambridge University Press.

ESA (2010). Mission Requirements Document GMES Sentinel-2. Retrieved December 15, 2010, from. http://esamultimedia.esa.int/docs/GMES/Sentinel-2_MRD.pdf

European Commission. (n.d.). GEOSS: Policy relevance, Future Challenges, EU contribution to GEOSS, Relevant documentation. *GEOSS*. Retrieved from http://ec.europa.eu/research/environment/index_en.cfm?section=geo&pg=geoss

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, *53*(4), 343–367.

Fang, H., Liang, S., Hoogenboom, G., Teasdale, J., & Cavigelli, M. (2008). Corn-yield estimation through assimilation of remotely sensed data into the CSM-CERES-Maize model. *International Journal of Remote Sensing*, *29*(10), 3011, doi:10.1080/01431160701408386.

Fang, H., Liang, S., Townshend, J. R., & Dickinson, R. E. (2008). Spatially and temporally continuous LAI data sets based on an integrated filtering method: Examples from North America. *Remote Sensing of Environment*, *112*(1), 75–93, doi:10.1016/j.rse.2006.07.026.

Féret, J. B., François, C., Asner, G. P., Gitelson, A. A., Martin, R. E., Bidel, L. P., et al. (2008). PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sensing of Environment*, *112*(6), 3030–3043.

Fung, A. K., & Chen, K. (2010). *Microwave scattering and emission models for users* (1st ed.). : Artech House Publishers.

Garrigues, S., Lacaze, R., Baret, F., Morisette, J. T., Weiss, M., Nickeson, J. E., et al. (2008). Validation and intercomparison of global Leaf Area Index products derived from remote sensing data. *Journal of Geophysical Research*, *113*(G2), G02028.

Ghil, M., & Malanotte-Rizzoli, P. (1991). Data assimilation in meteorology and oceanography. *Advances in Geophysics*, *33*, 141–266.

Giering, R., & Kaminski, T. (1998). Recipes for adjoint code construction. *ACM Transactions on Mathematical Software (TOMS)*, *24*(4), 437–474.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. .

Gobron, N., Belward, A., Pinty, B., & Knorr, W. (2010). Monitoring biosphere vegetation 1998–2009. *Geophysical Research Letters*, *37*(15), L15402.

Gobron, N., Pinty, B., Verstraete, M. M., & Govaerts, Y. (1997). A semidiscrete model for the scattering of light by vegetation. *Journal of Geophysical Research*, *102*(D8), 9431–9446.

Gobron, N., Pinty, B., Verstraete, M., & Widlowski, J. (2000). Advanced vegetation indices optimized for up-coming sensors: Design, performance, and applications. *Geoscience and Remote Sensing, IEEE Transactions on*, *38*(6), 2489–2505, doi:10.1109/36.885197.

Gobron, N., Pinty, B., Verstraete, M. M., & Widlowski, J. L. (2002). Advanced vegetation indices optimized for up-coming sensors: Design, performance, and applications. *Geoscience and Remote Sensing, IEEE Transactions on*, *38*(6), 2489–2505.

Goel, N. S. (1988). Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data. *Remote Sensing Reviews*, *4*(1), 1–212.

Goel, N. S., & Thompson, R. L. (2000). A snapshot of canopy reflectance models and a universal model for the radiation regime. *Remote Sensing Reviews*, *18*(2), 197–225.

Goward, S. N., Tucker, C. J., & Dye, D. G. (1985). North American vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer. *Plant Ecology*, *64*(1), 3–14.

Jacquemoud, S., & Baret, F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment*, *34*(2), 75–91.

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*(1), 1.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Kimes, D. S., Knyazikhin, Y., Privette, J. L., Abuelgasim, A. A., & Gao, F. (2000). Inversion methods for physically-based models. *Remote Sensing Reviews*, *18*(2), 381–439.

Knorr, W., Kaminski, T., Scholze, M., Gobron, N., Pinty, B., Giering, R., et al. (2010). Carbon cycle data assimilation with a generic phenology model. *Journal of Geophysical Research*, *115*(G4), doi:10.1029/2009JG001119.

Kuusk, A. (1995). A fast, invertible canopy reflectance model. *Remote Sensing of Environment*, *51*(3), 342–350.

Lyapustin, A., & Knyazikhin, Y. (2001). Green's function method for the radiative transfer problem I. Homogeneous non-Lambertian surface. *Applied Optics*, *40*, 3495–3501.

Lyapustin, A., Wang, Y., Martonchik, J., Privette, J. L., Holben, B., Slutsker, I., Sinyuk, A., & Smirnov, A. (2006). Local Analysis of MISR Surface BRF and Albedo Over GSFC and Mongu AERONET Sites. *IEEE Transactions on Geoscience and Remote Sensing*, *44*, 1707–1718, doi:10.1109/TGRS.2005.856678.

Lavergne, T., Kaminski, T., Pinty, B., Taberner, M., Gobron, N., Verstraete, M. M., et al. (2007). Application to MISR land products of an RPV model inversion package using adjoint and Hessian codes. *Remote Sensing of Environment*, *107*(1–2), 362–375.

Lewis, P., & Disney, M. (2007). Spectral invariants and scattering across multiple scales from within-leaf to canopy. *Remote Sensing of Environment*, *109*(2), 196–206.

Liang, S., & Strahler, A. H. (2002). An analytic BRDF model of canopy radiative transfer and its inversion. *Geoscience and Remote Sensing, IEEE Transactions on*, *31*(5), 1081–1092.

Lu, X., Liu, R., Liu, J., & Liang, S. (2007). Removal of noise by wavelet method to generate high quality temporal data of terrestrial MODIS products. *Photogrammetric Engineering and Remote Sensing*, *73*(10), 1129.

Lubansky, A. S., Yeow, Y. L., Leong, Y. -K., Wickramasinghe, S. R., & Han, B. (2006). A general method of computing the derivative of experimental data. *AIChE Journal*, *52*, 323–332, doi:10.1002/aic.10583.

McLaughlin, D. (2002). An integrated approach to hydrologic data assimilation: Interpolation, smoothing, and filtering. *Advances in Water Resources*, *25*(8–12), 1275–1286.

Myneni, R. B., Maggion, S., Iaquinta, J., Privette, J. L., Gobron, N., Pinty, B., et al. (1995). Optical remote sensing of vegetation: Modeling, caveats, and algorithms. *Remote Sensing of Environment*, *51*(1), 169–188, doi:10.1016/0034-4257(94)00073-V.

NASA. (n.d.). NASA - A-Train. Retrieved December 15, 2010, from http://atrain.nasa.gov/.

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., et al. (2003). Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *Science*, *300*(5625), 1560.

Olioso, A., Inoue, Y., Ortega-Farias, S., Demarty, J., Wigneron, J. P., Braud, I., et al. (2005). Future directions for advanced evapotranspiration modeling: Assimilation of remote sensing data into crop simulation models and SVAT models. *Irrigation and Drainage Systems*, *19*(3), 377–412.

Price, J. C. (1990). On the information content of soil reflectance spectra. *Remote Sensing of Environment*, *33*(2), 113–121.

Privette, J. L., Myneni, R. B., Tucker, C. J., & Emery, W. J. (1994). Invertibility of a 1-D discrete ordinates canopy reflectance model. *Remote Sensing of Environment*, *48*(1), 89–105.

Qin, J., Liang, S., Li, X., & Wang, J. (2008). Development of the adjoint model of a canopy radiative transfer model for sensitivity study and inversion of leaf area index. *Geoscience and Remote Sensing, IEEE Transactions on*, *46*(7), 2028–2037.

Qin, J., Liang, S., Liu, R., Zhang, H., & Hu, B. (2007). A weak-constraint-based data assimilation scheme for estimating surface turbulent fluxes. *Geoscience and Remote Sensing Letters, IEEE*, *4*(4), 649–653.

Qin, J., Liang, S., Yang, K., Kaihotsu, I., Liu, R., & Koike, T. (2009). Simultaneous estimation of both soil moisture and model parameters using particle filtering method through the assimilation of microwave signal. *Journal of Geophysical Research*, *114*(D15), D15103.

Quaife, T., & Lewis, P. (2010). Temporal constraints on linear BRDF model parameters. *Geoscience and Remote Sensing, IEEE Transactions on*, *48*(5), 2445–2450.

Quaife, T., Lewis, P., De Kauwe, M., Williams, M., Law, B. E., Disney, M., et al. (2008). Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter. *Remote Sensing of Environment*, *112*(4), 1347–1364.

Richardson, A. J., & Wiegand, C. L. (1977). Distinguishing vegetation from soil background information (by gray mapping of Landsat MSS data). *Photogrammetric Engineering and Remote Sensing*, *43*, 1541–1552.

Rochdi, N., & Fernandes, R. (2010). Systematic mapping of Leaf Area Index across Canada using 250-meter MODIS data. *Remote Sensing of Environment*, *114*(5), 1130–1135.

Rodgers, C. (2000). *Inverse methods for atmospheric sounding: Theory and practice*. World Scientific Publishing Company.

Roy, D. P., Jin, Y., Lewis, P. E., & Justice, C. O. (2005). Prototyping a global algorithm for systematic fire-affected area mapping using MODIS time series data. *Remote Sensing of Environment*, *97*(2), 137–162.

Slater, A. G., & Clark, M. P. (2009). *Snow data assimilation via an ensemble Kalman filter*.

Sobrino, J. A., Jiménez-Muñoz, J. C., & Verhoef, W. (2005). Canopy directional emissivity: Comparison between models. *Remote Sensing of Environment*, *99*(3), 304–314.

Stöckli, R., Rutishauser, T., Dragoni, D., O'Keefe, J., Thornton, P. E., Jolly, M., et al. (2008). Remote sensing data assimilation for a prognostic phenology model. *Journal of Geophysical Research*, *113*, doi:10.1029/2008JG000781 19 PP.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics.

Tha Paw, U., & others (1992). Development of models for thermal infrared radiation above and within plant canopies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *47*(2–3), 189–203.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150.

Twomey, S. (2002). *Introduction to the mathematics of inversion in remote sensing*. Courier Dover Publications.

Verhoef, W., & Bach, H. (2003). Simulation of hyperspectral and directional radiance images using coupled biophysical and atmospheric radiative transfer models. *Remote Sensing of Environment*, *87*(1), 23–41.

Vermote, E. F., El Saleous, N. Z., & Justice, C. O. (2002). Atmospheric correction of MODIS data in the visible to middle infrared: First results. *Remote Sensing of Environment*, *83*(1–2), 97–111.

Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics.

Weiss, M., Baret, F., Myneni, R. B., Pragnère, A., & Knyazikhin, Y. (2000). Investigation of a model inversion technique to estimate canopy biophysical variables from spectral and directional reflectance data. *Agronomie*, *20*(1), 3–22.

Widlowski, J. L., Taberner, M., Pinty, B., Bruniquel-Pinel, V., Disney, M., Fernandes, R., et al. (2007). Third radiation transfer model intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models. *Journal of Geophysical Research*, *112*(D9), D09111.

Xiao, Z., Liang, S., Wang, J., Song, J., & Wu, X. (2009). A temporally integrated inversion method for estimating leaf area index from MODIS data. *Geoscience and Remote Sensing, IEEE Transactions on*, *47*(8), 2536–2545.

Xiao, Z., Liang, S., Wang, J., Jiang, B., & Li, X. (2011). Real-time retrieval of Leaf Area Index from MODIS time series data. *Remote Sensing of Environment*, *115*(1), 97–106, doi:10.1016/j.rse.2010.08.009.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, *23*(4), 550–560.

Zupanski, D. (1997). A general weak constraint applicable to operational 4DVAR data assimilation systems. *Monthly Weather Review*, *125*, 2274–2292, doi:10.1175/1520-0493.