

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

P. Lewis\*<sup>1</sup>, J. Gomez-Dans<sup>1</sup>, T. Kaminski<sup>2</sup>, J. Settle<sup>3</sup>, T. Quaife<sup>4</sup>, N. Gobron<sup>5</sup>, J. Styles<sup>6</sup>, M. Berger<sup>7</sup>

1. Department of Geography, UCL, and National Centre for Earth Observation, Gower St., London, WC1E 6BT, UK.

2. FastOpt, Lerchenstr. 28a, D-20767 Hamburg, Germany.

3. National Centre for Earth Observation, University of Reading, Reading RG6 6AL, UK

4. College of Life and Environmental Sciences, University of Exeter and National Centre for Earth Observation, Peter Lanyon Building, Penryn, Cornwall, TR10 9EZ, UK.

5. 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.

6. Assimila Ltd., 1 Earley Gate, Reading RG6 6AT, UK.

7. ESA ESRIN, Science Strategy, Coordination and Planning Office (EOP-SA), Via Galileo Galilei, Casella Postale 64, 00044 Frascati (RM), Italy.

**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 track uncertainties. We look forward in the future to operational data assimilation schemes to track land surface processes and exploit multiple types of observation. 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

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

36

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

44

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

51

## 52 **1 Introduction**

### 53 **1.1 Background**

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

60

61 Early studies in terrestrial vegetation monitoring from EO (Richardson and Wiegand, 1977; Tucker,  
62 1979) found that simple transformations of multispectral measurements at red and near infrared  
63 wavelengths gave a signal that was responsive to the relative amount of green biomass and that  
64 could be used to track vegetation dynamics (Goward et al., 1985). The attractions of such  
65 ‘Vegetation Indices’ (VIs) are obvious: they are visually impressive as spatial and temporal datasets;  
66 they are simple to produce and provide a single quantity to interpret; they compensate for some of  
67 the extraneous factors that can otherwise complicate lower level EO signals; and they can often  
68 provide effective information for time series analyses, where the timing, rather than the magnitude  
69 of events is of importance (e.g. vegetation phenology). Further, such indices can be directly targeted  
70 at particular functional or physical vegetation properties, such as the fraction of absorbed  
71 photosynthetically active radiation (fAPAR) or Leaf Area Index (LAI), by design (Gobron et al.,  
72 2002, 2010) or empirically (Rochdi and Fernandes, 2010). In the former case a calibration is  
73 achieved using a set of radiative transfer model runs over a range of conditions (Gobron et al.,  
74 2000). In the latter, extensive ground-based measurements must be made (Chen et al., 2002) and the  
75 form of the relationship with a particular VI assumed. Such efforts are fast to process and often  
76 effective, especially for near-real-time survey. They have a range of known failings (Baret and  
77 Guyot, 1991), but some of these, such as dependence on the angular conditions of data acquisition

78 can be reduced by treating the data to normalise for such effects (e.g. Rochdi and Fernandes, 2010).  
79 Ultimately though, however much care is taken to treat such effects, methods assuming such fixed  
80 mappings from VIs with ‘statistical’ models are open to many criticisms, some of the more  
81 significant of which could be considered: (i) they fail to make full use of the information content of  
82 the observational data; (ii) they (often) fail to make use of our understanding of the physics of the  
83 situation; (iii) they need recalibration if conditions change (e.g. sensor band pass functions or scale  
84 of observation); (iv) they tend not to treat uncertainty in the mapped product in any rigorous way  
85 (mostly, they fail to consider this at all). This is a judgement call.

86

87 An alternative stratagem has been to build mathematical models of the physics of radiation  
88 interactions with vegetation canopies and the intervening atmosphere, phrased as functions of  
89 ‘control’ variables (polarisation, wavebands, viewing and illumination angles etc.) and (bio) physical  
90 parameters or ‘state variables’ (LAI, leaf chlorophyll concentration etc. for the canopy, and aerosol  
91 optical depth, ozone concentration etc. for the atmosphere), and to use these to attempt to interpret  
92 the satellite signal. We may call these radiative transfer (RT) models. To tie in with discussions  
93 below and to provide consistency with the data assimilation literature, such models are called here  
94 ‘observation operators’ (denoted  $H(x)$ ) in that they map from the state variable vector  $x$  to the EO  
95 signal (as a vector)  $R$  for a given set of control variables, so the modelled signal vector  $R = H(x)$ .

96 The ‘remote sensing inverse problem’ then is to obtain an estimate of some function of  $x$ ,  $F(x)$   
97 from measurements  $R$ . How this may be achieved is discussed in more detail below.

98

99 Much effort has been devoted to producing information from EO data about specific biophysical  
100 quantities that are relevant to science and society. A major focus of this has been to attempt to  
101 provide estimates of (green) LAI. Garrigues et al. (2008) consider four representative EO-derived  
102 global LAI products, with core spatial resolutions of 1 km or coarser, that use what might be

103 considered state of the art methods for multi-year dataset generation. The reader is referred to that  
104 paper for detailed information on the products, a product inter-comparison and validation against  
105 independent ground measurements. The temporal resolution of the products varies from 8 days to 1  
106 month. Three of the products (ECOCLIMAP, GLOBCARBON (V1), and CCRS) are derived from  
107 assumed VI relationships with LAI. A fourth (MODIS (C4)) uses such a relationship for a backup  
108 algorithm. Three of the products (GLOBCARBON, CYCLOPES (V3.1) and MODIS) make use of  
109 RT models in attempting to estimate the LAI. In the case of GLOBCARBON the RT model is used  
110 to calibrate the VI-LAI relationship. For MODIS a look up table derived from the RT model is used  
111 to map red and near infrared (NIR) bidirectional reflectance data to LAI, and for CYCLOPES a  
112 neural network derived from an RT model is used for the mapping from red, NIR and shortwave  
113 infrared (SWIR) portions of the electromagnetic spectrum. A feature of these uses of RT models is  
114 that they can map many channels of input data to one (or many) outputs. The one-to-one mapping  
115 used in VI design and/or calibration is then just the simplest case of this more general RT approach.

116

117 A major new effort in satellite data provision is the GMES (Global Monitoring for Environment and  
118 Security; [www.gmes.info](http://www.gmes.info)) programme (Council of the European Union, 2010). It is an EU initiative  
119 set up to provide timely information on key environmental variables for policy makers and public  
120 authorities, and is intended to be a major EU contribution to understanding and managing climate  
121 change. Six thematic areas are being developed: marine, land, atmosphere, emergency and security  
122 and climate change. The land monitoring service is provided via the GEOLAND2 project  
123 ([www.gmes-geoland.info](http://www.gmes-geoland.info)), which oversees the generation of products derived from satellite data,  
124 providing information on a wide range of variables including LAI. GMES is a European  
125 contribution to GEOSS, the Global Earth Observing System of Systems (European Commission,  
126 n.d.). The Sentinels are a series of satellites being developed by the European Space Agency that are  
127 specifically designed to address the space observation requirements of GMES. There are five  
128 Sentinel missions, each of which will consist of a pair of satellites (for details see Aschbacher et al.,

129 2012 and dedicated Sentinel mission papers, all this RSE issue). This paper is primarily concerned  
130 with methods for the retrieval of biophysical parameters of terrestrial ecosystems, including LAI,  
131 from instruments at arbitrary spatial resolutions, sun-sensor geometries and optical wavelengths.  
132 Consequently the techniques described here are directly relevant to Sentinels 2 and 3 missions.  
133 Sentinel 2 has a medium resolution multispectral imager (MSI) in the optical domain with 4 bands  
134 at a 10m resolution, 6 bands at 20m and 3 bands at 60m. These 13 spectral channels (Table 1) are  
135 distributed in the visible and near infrared and shortwave infrared regions. The Ocean Land Color  
136 Instrument (OLCI) instrument on board the Sentinel 3 platform is a coarser (circa 500m) resolution  
137 instrument, similar to MERIS that is designed for global monitoring applications. In principle the  
138 system described in this paper could also be extended to other wavelength domains and  
139 consequently be used to integrate data from the entire suite of EO missions.

140

141 TABLE 1 ABOUT HERE

142

143 An additional context for this paper is the growing interest in the application of wider constellations  
144 of satellites for environmental and disaster monitoring. A manifestation of this is the NASA A-train  
145 (NASA, 2010), which is a formation of complementary satellites and sensors taking observations at  
146 close to the same time. Other examples include relatively low cost satellites and instruments with a  
147 suite of similar instruments flying in formation to provide global daily viewing opportunities at  
148 mid-resolution (10-30m), for example the Disaster Monitoring Constellation (DMC) (DMCII,  
149 2010). The concept can potentially be applied to more heterogeneous systems, such as the ‘virtual  
150 constellation’ for Land Surface Imaging (LSI) concept promoted by the Committee on Earth  
151 Observation Satellites (CEOS) to optimise benefits from land remote sensing systems (CEOS,  
152 2011a). There are clear benefits for monitoring frequency if data from a wider range of sensors are  
153 available, but the more heterogeneous the set of sensors (in terms of spatial resolution and  
154 wavelength domains) the more important it is to formalize appropriate methods to optimally merge

155 information from these sources.

156

## 157 **1.2 Optimal estimation**

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

168

$$169 \rho(x) = \exp(-J(x))$$

170

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

173

$$174 J(x) = \sum_i \hat{a}_i J_i(x) \tag{1}$$

175

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

177

178 Much of the earlier literature on estimating  $x$  for vegetation monitoring from a physical basis  
179 concentrated on exploring options in numerical minimisation approaches (see e.g. the review by

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

205



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

231

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

237

### 238 **1.3 Data Assimilation**

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

247

248 Early examples of data assimilation systems are those used to improve short-range weather  
249 predictions from meteorological models (Ghil and Malanotte-Rizzoli, 1991). In these systems the  
250 number of state variables is typically huge, often greater than  $10^6$ , because of the large number of  
251 interconnected sub-domains used to represent the atmosphere in a 3D grid. The number of  
252 observations available is typically several orders of magnitude less than this, and in consequence  
253 the problem is ill-posed. However, including a constraint that the final solution should not diverge  
254 too far from an *a priori* estimate (typically supplied by a previous model run) tends to result in a  
255 tractable solution. The schemes used for these problems are referred to as 'variational', being based  
256 in the field of mathematics dealing with the calculus of variations, and are closely related to the DA  
257 system described in this paper. A 'strong constraint' variational DA system assumes that the

258 underlying process model prescribing the state vector evolution is correct (i.e. there is a model  
259 trajectory that matches the observations). In this case it is generally only the initial state of the  
260 system that is estimated by the DA procedure, but this approach can also be used to calibrate models  
261 (i.e., to optimise estimates of model process parameters) (Knorr et al., 2010). If the state vector is  
262 allowed to deviate from the model predictions then this is referred to as a 'weak constraint' DA  
263 system (Zupanski, 1997). It is this latter type that is used here and is discussed more completely in  
264 later sections.

265

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

272

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

284

285 The Earth Observation Land Data Assimilation System (EO-LDAS) study funded by ESA aims at  
286 supporting the generation of a generic land data assimilation system by using the full information  
287 content provided by observations from satellite constellations. Such a system, in eventual  
288 operational form, is intended primarily to improve the quality and consistency of land surface  
289 products generated from multi-sensor EO data. The project is focussed on developing a generic  
290 scheme and software prototyping for use with medium to mid spatial resolution (in the range 10m –  
291 500m) optical data. The principal design concept is to allow integration of data from different  
292 satellites observing the surface of the earth at different sun-sensor geometries, wavebands and  
293 spatial scales, such as that supplied by Sentinels 2 and 3, in a physically consistent manner, and to  
294 provide information on the state of the surfaces with well-quantified estimates of uncertainty. It also  
295 demonstrates the idea that predictions based on data from one sensor can be made from a DA  
296 system driven by observations from another, a concept that could potentially be used to aid  
297 vicarious sensor calibration.

298

## 299 **2 The EO-LDAS prototype**

### 300 **2.1 The EO-LDAS Scheme**

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

309

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

321

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

329 Thus, in EO-LDAS, equation 1 becomes:

330

331 
$$J(x) = J_{obs}(x) + J_{prior}(x) + J_{model}(x)$$

332

333 Each of these constraints has associated with it an error model. In the following sections, we

---

<sup>1</sup> <http://www2.geog.ucl.ac.uk/~plewis/eoldas/>

334 describe the set of constraints and the DA algorithm. We stress that in the text below, we use the  
335 symbol  $x$  to refer to the set of state variables that we wish to estimate. In EO-LDAS this essentially  
336 means a representation of the state at each sample points in time (and/or space) that we consider.  
337 So, for example if we were trying to estimate Leaf Area Index and leaf Chlorophyll content at one  
338 location for every day of the year, we would have a state vector with 361x2 elements. In addition,  
339 EO-LDAS has the capacity to augment this state vector with ‘static’ state representations (some  
340 term affecting one or more of the constraints that we wish to be considered constant in space/time).

341

## 342 **2.2 Observational Constraint**

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

346

$$347 \quad J_{obs}(x) = \frac{1}{2} (R - H(x))^T C_{obs}^{-1} (R - H(x)) \quad (2)$$

348

349 where  $^T$  denotes the transpose operator. This is the penalisation associated with differences between  
350 the predicted and observed reflectance values. The covariance matrix  $C_{obs}$  describes the uncertainty  
351 in the observations (and also formally, in the observation operator). As noted, the purpose of the  
352 observation operator  $H(x)$  is to translate information from the state space to that of the  
353 observations, and is in practice a radiative transfer model. For ease of implementation (mainly  
354 involving spectral sampling issues), when different sensor types are used in EO-LDAS, a set of  
355  $J_{obs}(x)$  terms is developed, with one for each sensor type.

356

357 There have been many attempts to create observation operators  $H(x)$ , varying in complexity,  
358 accuracy and computational cost. Goel (1988) provides a review of most of the concepts for

359 radiative transfer model developed for reflectance from vegetation canopies at optical wavelengths  
360 (see also (Goel and Thompson, 2000), with (Tha Paw U, 1992) covering related materials for  
361 thermal emitted radiation and (Fung and Chen, 2010) for the microwave domain. Some updates and  
362 model intercomparisons are provided by Sobrino et al. (2005) (thermal) and Widlowski et al.,  
363 (2007) (optical). The focus in this paper, and in the prototype EO-LDAS is on the use of optical  
364 sensor data, but the approach outlined here is easily adapted for use in other wavelength domains.

365

366 In a similar way, atmospheric properties, such as aerosol optical depth and water vapour content,  
367 need to be accounted to obtain accurate estimates of surface properties. This can be achieved by  
368 coupling the surface model with an atmospheric model, and solving for both the surface and  
369 atmosphere parameters simultaneously (Verhoef and Bach, 2003). Some (probably most)  
370 approaches to surface interpretation use surface reflectance that has already been ‘corrected’ for  
371 atmospheric effects (Vermote et al., 2002), but a full decoupling of the problem, at optical  
372 wavelengths at least, cannot be achieved without knowledge of the surface Bidirectional  
373 Reflectance Distribution Function (BRDF) (Lyapustin and Knyazikhin, 2001; Lyapustin et al. 2006)  
374 (or at least a normalised form of this) (Vermote et al., 1997).

375

376 The observation operator we use in this paper is developed from the original semi-discrete model of  
377 Gobron et al., (1997). It has a state vector describing canopy architecture and three spectral terms,  
378 although these are all defined as functions of other parameters as described below (Table 2). The  
379 soil reflectance is assumed Lambertian in the model, although it could be adapted to incorporate a  
380 soil directional reflectance model. As stated here then, the (canopy-soil) model estimates the  
381 directional reflectance factor at a set of viewing and illumination angles for a given narrow  
382 waveband. Since the model must be capable of predicting the reflectance at arbitrary (solar  
383 reflective) wavelengths, spectral models are incorporated in the code to predict the soil  
384 (Lambertian) reflectance and leaf (bi-Lambertian) reflectance and transmittance. Since model

385 derivatives are required, we use for simplicity here: (i) the linear soil reflectance model of Price  
386 (1990); and (ii) an approximation to the PROSPECT leaf reflectance/transmittance model of (Féret  
387 et al., 2008), being a minor modification of the model of (Jacquemoud and Baret, 1990). The  
388 approximation was developed for possible processing speed enhancements, but is identical in form  
389 to PROSPECT if the parameter N (table 2) is 1, and very close to the original model over the range  
390 of N 0.8 to 2.5.

391

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

397

398 TABLE 2 ABOUT HERE

399 TABLE 3 ABOUT HERE

400

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

408

409 Differentiated versions of the observational cost are required to enable the use of efficient gradient  
410 descent minimisation routines, so we can benefit from access to  $J_{obs}^c(x)$ , the derivative of  $J_{obs}(x)$



411 with respect to  $x$ . This is:

412

$$413 \quad J_{obs}^{\mathbb{C}}(x) = -H^{\mathbb{C}}(x)^T C_{obs}^{-1} (R - H(x)) \quad (3)$$

414

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

423

424 We obtain an estimate of the posterior uncertainty through consideration of the curvature at the  
425 global minimum in state space. This is provided by the inverse of the sum of the constraint  
426 Hessians, the Hessian for this constraint being  $J_{obs}^{\mathbb{C}}(x)$ :

427

$$428 \quad J_{obs}^{\mathbb{C}}(x) = H^{\mathbb{C}}(x)^T C_{obs}^{-1} H^{\mathbb{C}}(x) - H^{\mathbb{C}}(x)^T C_{obs}^{-1} (R - H(x)) \quad (4)$$

429

430 Although it should be possible to develop a Hessian code in much the same way as done for the first  
431 derivative, that has not yet been done within EOLDAS, so a linear approximation to the Hessian is  
432 achieved, using finite differences. As we will see below, the algorithm used to perform the  
433 optimisation is iterative, but the potentially high cost of using finite differences for the Hessian is  
434 unimportant in this sense, as it only has a role in estimating the posterior uncertainties.

435

### 436 2.3 Process model constraint

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

$$445 \quad M(x) = Ax + b \quad (5)$$

446  
447 where  $A$  and  $b$  are a matrix and vector respectively. One advantage of designing the prototype  
448 system in this manner is that it provides a flexible framework for changing the underlying model.  
449 Unlike in a sequential system, this formulation directly allows for any model state vector element to  
450 be linked to any other, since  $x$  here contains the state representation at all sample times (spaces), so  
451 different time/space scales can be readily incorporated. The cost function associated with this  
452 process model then is:

$$454 \quad J_{model}(x) = \frac{1}{2}(x - M(x))^T C_{model}^{-1} (x - M(x)) = \frac{1}{2}((I - A)x - b)^T C_{model}^{-1} ((I - A)x - b) \quad (6)$$

455  
456 where  $I$  is the identity operator.  $J_{model}(x)$  is the cost incurred by departure of the model state from  
457 that predicted by an underlying process model. An interpretation of  $A$  is as the model derivative.  
458 The model uncertainty matrix  $C_{model}$  therefore expresses the uncertainty in this derivative, including  
459 any inherent uncertainty in the process model. In such a case, it might often be pragmatic to specify  
460 only diagonal terms in  $C_{model}$  as further details of model structure are often difficult to obtain. In

461 any case, we can see that EO-LDAS could be interfaced to a process model such as the Carbon Flux  
462 model DALEC used by Quaife et al. (2008) or any other for which the derivative might be obtained  
463 (e.g. using AD) by augmenting the state vector  $x$  by any terms that we might wish to drive the  
464 model.

465

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

475

476 We can for example consider the benefits of approaches such as Twomey-Tikhonov regularisation  
477 or variations around this theme (Rodgers, 2000). Examples of this that we explore further below are  
478 first and second order difference constraints. In essence these improve the conditioning of the  
479 inverse problem by smoothing or regularising the solution, which comes about because they  
480 constrain derivatives (first or second order here) to be zero. In a weak constraint DA system such as  
481 that used here, the model is not strictly enforced (this would be clearly undesirable in these  
482 derivative constraints) but rather the degree of smoothness in the outcome is traded off against the  
483 other factors in  $J(x)$  through the model uncertainty matrix. In other words, the cost function will  
484 penalise temporal trajectories of parameters that are not flat, but this is ‘balanced’ with a goodness  
485 of fit to the observations and departure from the prior estimate. In practice this constrains the  
486 solution toward a smooth evolution by minimising the high frequency components of the temporal

487 parameter trajectory. A similar approach has been taken by Quaife and Lewis (2010) for linear  
 488 observation operators. Viewing this form of solution as a combination of state variable estimation  
 489 and filtering, we note that the filter characteristics are controlled by the nature of matrices  $A$  and  
 490  $C_{model}$ , the former controlling the cut-off frequency of the filter and the latter, if simply diagonal,  
 491 controlling the degree of dampening of the unwanted high frequencies. In this context, we can  
 492 consider  $b$  a bias term, which we set to zero. In this case:

493

$$494 \quad J_{model}(x) = \frac{1}{2}(Dx)^T C_{model}^{-1}(Dx)$$

495

496 where  $D = (I - A)$ . The derivatives of this are:  $J_{model}^{\mathbb{C}}(x) = D^T C_{model}^{-1} Dx$  and  $J_{model}^{\mathbb{C}}(x) = D^T C_{model}^{-1} D$ . To  
 497 achieve Twomey-Tikhonov regularisation then, which we view as an empirical process model,  $D$   
 498 here becomes simply a ( $N^{\text{th}}$  order) differential operator (Quaife and Lewis, 2010). In many  
 499 situations, we must assume the uncertainty in this empirical constraint unknown. The minimum  
 500 error model then is a constant value for which we can use a scalar term  $g$ :

501

$$502 \quad J_{model}(x) = \frac{g^2}{2} x^T (D^T D) x \tag{7}$$

503

504 We can interpret  $g$  as a ‘smoothness term (or  $g^{-1}$  as a roughness term) that controls the weighting  
 505 of the derivative (model) constraint with respect to the other constraints. It is worthwhile at this  
 506 point trying to relate this back to the discussions on process models. This is most readily achieved  
 507 by considering a first order derivative constraint. If applied at lag 1 day for a temporal constraint,  
 508 we can interpret this as an expectation that the state vector tomorrow will be the same as today (i.e.  
 509 the derivative is zero). If we want to relate this to equivalent sequential methods, we can say that  
 510 this is a zero-order process model. The term  $g^{-1}$  then can be interpreted as uncertainty (phrased as

511 standard deviation) in this model, or alternatively as the growth in uncertainty over a one day  
512 period. Similar interpretations apply for other derivative constraints: a second order derivative  
513 constraint is equivalent to a first order process model. Equation 5 then is a viable empirical process  
514 model constraint, but we have yet to tackle the fact that the smoothness  $g$  is unknown. We also note  
515 that if we use a scalar for  $g$ , we are assuming the same smoothness for all state variables at all times  
516 (places).

517

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

525

526 It is generally found (e.g. Twomey (2002)) that quite a broad range of model uncertainty  
527 (smoothness) estimates can provide an acceptable solution, so we do not expect the results to be  
528 overly-sensitive to the choice of this ‘hyper-parameter’. We could make a rough guess at the model  
529 uncertainty, but that is likely to be unsatisfactory in the general case. If we under-estimate it by too  
530 much, we can over-dampen most of the state vector. Equally, if we greatly over-estimate the model  
531 uncertainty, the impact of the temporal constraints is minimal: in the extreme, an infinite model  
532 uncertainty (zero smoothness) leads to a solution without model constraint. Whilst there are several  
533 strategies that can be employed to estimate the model uncertainties (hyper-parameters), perhaps the  
534 most fruitful in the context of EO-LDAS is running a cross-validation exercise. The idea is that an  
535 independent dataset is used to test the robustness of the solution for a particular value of the hyper-  
536 parameters. An optimal estimate of the hyper-parameters (or distribution thereof) can be obtained

537 by minimising a cost function with the independent observations. This can be achieved with a sub-  
 538 set of observations to test a solution obtained from the rest of the dataset, a strategy that when  
 539 repeated over different subsets becomes known as generalised cross validation becomes known as  
 540 generalised cross validation (Wahba, 1990, Eilers, 2003 and Lubansky et al., 2006).. Alternatively,  
 541 we might use data from an independent sensor, although accurate absolute calibration between the  
 542 sensors is needed for that.

543

## 544 2.4 Prior Constraint

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

548

$$549 \quad J_{prior}(x) = \frac{1}{2} (x - x_{prior})^T C_{prior}^{-1} (x - x_{prior}) \quad (8)$$

550

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

558

## 559 2.5 DA algorithm

560 The various constraints discussed above provide the cost function in equation 1 through their

561 summation. This also applies to the derivatives  $\mathcal{J}(x)$  and  $\mathcal{C}(x)$ . The cost function  $J(x)$  is  
562 minimised using a gradient descent method (i.e. using  $\mathcal{J}(x)$ ). Bounds are applied as a final  
563 constraint to the solution, to ensure that the state vector remains within physical limits. These can be  
564 altered by the user for any particular run of the system via a configuration file or command line  
565 interface. In the EO-LDAS prototype we use the limited memory Broyden-Fletcher-Goldfarb-  
566 Shanno (L-BFGS-B) algorithm described in (Byrd et al., 1995; Zhu et al., 1997). In principle,  
567 however, a number of different gradient descent algorithms could be used. The L-BFGS-B was  
568 selected for its efficient memory handling for high dimensional problems and the fact that it can  
569 optimise over a bounded domain, which is appropriate for this problem.

570

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

576

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

585

586 The DA system developed here can be viewed an extension of the methodologies that have been

587 applied to inverting radiative transfer models by minimising a cost function. The addition of a linear  
588 dynamic model therefore only adds a handful of extra parameters to the problem (namely, the nature  
589 of the dynamic model itself and the associated covariance matrix,  $C_{model}$ , which may be simply  
590 diagonal). This is in a marked contrast with similar methodologies that either use a long time series  
591 of data for inverting one single parameter (in the case of inverting LAI as in (Fang et al. 2008b;  
592 Xiao et al. 2009; Xiao et al. 2011)). The temporal smoothness constraint is in itself an important  
593 feature, which is usually performed as a post-processing step after the parameter retrieval (Lu et al.  
594 2007).

595

### 596 **3. Experimentation**

597 We present a series of experiments to demonstrate the operation of the EO-LDAS prototype and to  
598 explore the sorts of capabilities such a system could provide with data from the Sentinel-2 MSI  
599 sensor (see table 1a for waveband information for Sentinel-2 MSI). The experiments use synthetic  
600 data for observations i.e. are derived from running the observation operator for a given state vector  
601 for what we suppose to be typical Sentinel-2 scenarios over one calendar year. We simulate top hat  
602 function bandpass functions (1 nm sampling) according to the information in Table 1 (see also  
603 Drusch et al., 2012, this issue). The main aim of the experiments is to determine the improvement,  
604 in terms of reduced uncertainty, in biophysical parameter estimation that might be obtained by  
605 applying the EO-LDAS prototype for such scenarios. A subsidiary aim is to demonstrate the  
606 capability of the DA system to make predictions of data from a sensor not used in the DA process.  
607 Here, we do this by using the state vector estimates derived from the DA with synthetic MSI data,  
608 and make predictions of what a SPOT-5-like instrument would view (described below). These data  
609 are used in a cross-validation exercise within the experiments.

610

#### 611 **3.1. Experimental setup**

612 In these experiments, we control the time trajectory of a subset of model parameters according to



613 the functions given in Table 4, where  $t$  is the relative day of year (DOY) i.e. DOY normalised by  
614 365. All other parameters take their default values given in Table 2. The functions for LAI and  
615 chlorophyll broadly mimic typical trajectories of these terms for crops: for LAI, a flat initial period,  
616 followed by a rise to maximum LAI and then a symmetric decrease; for Chlorophyll, a linear rise  
617 and decrease. The more arbitrary functions used for the soil brightness term  $s_1$  we include to mimic  
618 rather broad variations over the year that might be supposed to be responses to soil moisture. A  
619 similar function is used here for leaf water, with a time lag of 36.5 days. The quite large variation of  
620 these two latter terms is intended primarily to allow the operation of the data assimilation scheme to  
621 be explored over a wide range of conditions, rather than to too closely mimic some particular  
622 situation. In that context, the rather large time lag between soil brightness variation and leaf water  
623 content is unrealistic, but a larger phase between these terms should test the system to a greater  
624 extent than having all parameters following similar trajectories. Although the full set of state vector  
625 elements is 13 for each time sample, we attempt to retrieve only the 6 elements (no 1, 4, 6, 7, 8, 9 in  
626 table 2) (per time sample) that we vary in these experiments, i.e. we assume the remaining elements  
627 fixed and known. This is partly to reduce the computational time required for the DA and more  
628 broadly because we believe it is sufficient to demonstrate the principles underlying the DA method.  
629 It is quite feasible to permit an estimation of 12 of the 13 elements (not the categorical variable  
630 directly through this method) but this is not the purpose of this exercise, and (arbitrary) variations in  
631 these additional terms would need to be defined to achieve this.

632

633 TABLE 4 ABOUT HERE

634

635 To approximate the Sentinel-2 MSI acquisition geometry (ESA, 2010), we assume one sample  
636 every 5 days (73 samples over the year), with a solar zenith angle corresponding to 10:30 local time  
637 at  $50^\circ$  N, random relative azimuth and random view zenith between  $0^\circ$  and  $15^\circ$ . Whilst these  
638 parameters do not provide a precise prediction of the likely MSI sampling and geometries, they are

639 close enough to develop an understanding of the likely behaviour of the data. The random azimuth,  
640 for example, is clearly in error, but since the view zenith angle is so restricted, this will have very  
641 little impact; the local time at 50° N will in reality be slightly later than the nominal equatorial  
642 crossing time used here, but the details of the solar zenith angle are less important here than  
643 inducing a typical variation over the year (32° to 76° here). The simulation of one sample every 5  
644 days mimics close to the maximum sampling achievable by MSI on 2 Sentinel platforms.

645

646 Synthetic observations were also generated for a SPOT5 HRG-like instrument. This sensor has four  
647 wavebands (500-590, 610-680, 790-890 and 1530-1750 nm) (CEOS, 2011b). We have assumed a  
648 revisit period of 13 days (to be out of sync with the synthetic MSI observations), although the  
649 differences are only up to two days from the MSI observations. The view zenith angle was limited  
650 to +/-25 degrees from nadir, with a random azimuth and a local overpass time of 10.30. In total, 28  
651 observations were available in this dataset.

652

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

665 to treat only the surface (canopy-soil) elements of the observation operator.

666

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

673

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

686

687 We have initially tested the system without observational noise and confirm that the scheme  
688 retrieves the truth to within the bounds implied by the convergence criteria and machine precision.  
689 Processing time for a single set of 73 time samples with MSI spectral sampling, solving for 6 state  
690 vector elements for each day of the year (2190 in total), is currently several hours on a 3 GHz Intel

691 processor on a single core, but this is partially due to very stringent convergence criteria used whilst  
692 testing the code and partially because this prototype implementation requires some significant  
693 efforts in computer code optimisation.

694

695 In all experiments, we set the prior estimate of the state vector to the values shown in table 2, with  
696 very large diagonal uncertainty terms (8). This effectively removes the prior constraint from  
697 consideration in these experiments, as we wish to conduct experiments based only on model and  
698 observational constraints here.

699

700 In the following sections, we examine the result of applying the weak constraint variational data  
701 assimilation approach described above to the synthetic dataset. For all cases, we assume that the  
702 uncertainty in the observations is known and that it is Gaussian and uncorrelated between  
703 wavebands and between dates. In the first case (3.2), we solve for state vector estimates assuming  
704 no dynamic model constraint other than the weak prior (standard deviation 8). This acts as a  
705 baseline for further experiments. In the second case (3.3) we assume that model uncertainty  $g$  is  
706 unknown and attempt to solve for it and the state vector for each day in the year with a form of  
707 cross-validation exercise using the SPOT-5 HRV synthetic observations. The ‘true’ values of  $g$  for  
708 individual state vector elements are shown in Table 5. The DA is performed with the ‘complete’ (i.e.  
709 5-day sampling MSI) dataset in that case. Finally, we repeat that experiment for the ‘cloudy’ dataset  
710 (3.4).

711

712 TABLE 5 ABOUT HERE

713

714 Graphical results (figures 1-2) are presented as untransformed biophysical variables (i.e. LAI,  $C_{ab}$ ,  
715  $C_w$ ,  $C_{dm}$ ,  $N$  and  $s_1$ ), showing: the ‘true’ (‘original’) state vector (dashed line); circles and error bars  
716 (shaded region) shows mean and 95% credible interval bounds (at plus/minus 1.96 standard

717 deviations). We will term 1.96 standard deviations ‘uncertainty’ for the remainder of the paper,  
718 unless the statement is otherwise qualified. The uncertainty bounds are slightly larger for the upper  
719 limits than for the lower limits (other than for  $N$  and  $s_1$ ) due to the nature of the transformations  
720 used in the approximate linearization (table 3). Tabular results for the experiments (tables 9-12) are  
721 expressed in transformed parameter space, as that is the space in which the state vector is inferred  
722 and in which the Gaussian statistics derived are most natural.

723

### 724 **3.2 Baseline estimates**

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

728

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

734

735 FIGURE 1 ABOUT HERE

736

737 In figure 1, the column titled ‘single obs inversion’ shows the results of this state vector estimate for  
738 the six parameters that are varied, transformed back to their biophysical meanings (through the  
739 inverse of the functions in table 3). The sub-plots rows show results for the observation operator  
740 parameters LAI,  $C_{ab}$ ,  $C_w$ ,  $C_{dm}$ ,  $N$  and  $s_1$  respectively. The uncertainty (average credible interval)  
741 associated with each (transformed) parameter for the baseline experiment is given in table 6 (‘single  
742 obs.’). Relating these uncertainties to the parameter ranges (table 4), we note that they are around

743 5% for TLAI, 10% for  $s_1$  and  $TC_{ab}$  respectively for dates where there are observations, more than  
744 20% for (transformed) leaf water and dry matter content and around 33% for N. We can suppose  
745 these then to be typical uncertainty values for MSI sampling (with the assumed noise  
746 characteristics). The cross correlation associated with these, illustrated in table 7 are highly variable  
747 from one sample to the next. The median values given show quite strong negative correlations  
748 between TLAI and  $TC_{ab}$  and  $TC_w$  but positive correlations with  $s_1$  and  $TC_{dm}$ . The median  $s_1$  shows  
749 negative correlations with all terms other than TLAI. Despite the fact that the average transformed  
750 LAI uncertainty is only around 5%, we can see if figure 1 that both the error and uncertainty can be  
751 rather high. Around peak LAI, results from individual samples vary by around LAI 2.5 and there is  
752 a general tendency to underestimate. The general trends of  $C_{ab}$  and  $C_w$  are discernable, but there is  
753 large variation and large uncertainty. The terms that are supposed to be constant here,  $C_{dm}$  and N  
754 depart significantly from their true state and the negative correlation is evident in the state  
755 trajectories around the central part of the year.

756

757 TABLE 6 ABOUT HERE

758 TABLE 7 ABOUT HERE

759

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

766

767 In spite of these quite high levels of uncertainty (and correlation of uncertainty) for these estimates,  
768 there is clearly quite a strong correlation between the values of the state vector and its neighbours in

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

787

### 788 **3.3 DA: Complete scenario**

789 Here, we apply first order and second order derivative constraints to the solution, but we expect the  
790 results to be broadly similar. In both cases, we need only supply some estimate of the uncertainty  
791 associated with these constraints through the smoothness term  $g$  to achieve a regularised solution to  
792 the state vector estimate. These constraints are applied by incorporating a model that, in the absence  
793 of any observations, would set the first (second) derivatives of the state variables to zero. Assuming  
794 that we apply the same (strength of) constraint to the whole time series, we need to supply an

795 estimate of the mean squared first (second) difference in the parameter values (true values in table  
796 5). For the first (second) difference then, this can be thought of as an estimate of the uncertainty in a  
797 zero-order (first-order) process model over one time step as noted above.

798

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

809

810 FIGURE 2 ABOUT HERE

811

812 Figure 2 shows the error in cross-validation as a function of  $g$  for the model first- and second-order  
813 difference constraints for the complete case (black circles and squares respectively). There are clear  
814 minima for these functions, which provide estimates of the optimal model uncertainty (averaged  
815 over all terms). Also shown in the figure is a set of vertical lines that represent the theoretical value  
816 of the smoothness term for each of the state vector elements that vary over time (from table 5). For  
817 the first order constraint, the minimum of the cross validation function is  $g=150$  which is very close  
818 to the theoretical values. For the second order constraint the cross validation RMSE minimum at  
819  $g=530$  is rather less than the theoretical values. For both cases however, we observe a very broad  
820 minimum, so there is quite a large range of values of  $g$  that allow almost equally good prediction of



821 the synthetic cross validation HRG observations.

822

823 TABLE 8 ABOUT HERE

824

825 Tables 8 provides statistics on the uncertainty reduction, (the posterior uncertainty estimate from the  
826 DA relative to that after solving for each sample separately and assuming the prior uncertainty  
827 where there are no observations). The average improvement in uncertainty is 4.07 for the first order  
828 constraint and 2.73 for the second order difference constraint. This is very significant but it must be  
829 remembered that 4/5 of the samples in the ‘single obs’ solution have only the prior constraint and  
830 uncertainty. Examining only locations where observation lie (i.e. ignoring interpolation  
831 performance relative to the *a priori* estimate), we see the uncertainty reduction drop by nearly 50%  
832 in this case, down to 2.20 for the first order constraint and 1.30 for the second difference constraint.  
833 From those figures, we would suppose the first order constraint to be greatly superior to the second  
834 order constraint, but if we look at the plots in figure 1, the second order constraint results seem to  
835 have more reasonable uncertainty bounds than the other results. This is at least partially because the  
836 apparent uncertainty resulting from the DA is strongly dependent on the value of  $g$  used in the  
837 model constraint: the higher the value of  $g$ , the smoother will be the solution and the lower the  
838 estimate of uncertainty. The only check we have done on the *veracity* of the solution comes from  
839 the cross validation, which is an indirect check: in any non-synthetic experiment we rarely know the  
840 ‘truth’ to any great degree of certainty. Since we have a synthetic experiment here, we can however  
841 test how frequently the derived solution matches the (synthetic) truth within the claimed uncertainty  
842 bounds. One reasonable summary measure of this is the percentage of true values of state vector  
843 elements that lie within the 95% credible interval claimed by the DA results. These are shown in  
844 table 8. We can see that for the ‘single obs’ estimates (no regularisation), only around 64% of the  
845 state vector lies within the 95% credible interval claimed by the solution. The figure is as low as  
846 58% for  $TC_{dm}$ . We can suppose the average estimated uncertainty then to be only around 67% of

847 the true value, i.e. we should inflate the estimated uncertainty by a factor of around 1.5. This would  
848 apply equally to the results in table 6. We see almost the same value for the first order constraint,  
849 which suggests the reduction in uncertainty by a factor of 2.2 is likely true. For the second order  
850 difference constraint however, around 84% (table 8) of the sample lie within the uncertainty bounds,  
851 so here, a better estimate of the uncertainty reduction might be around 1.70 rather than the 1.30  
852 reported. This apparent under-reporting of the uncertainty is worthy of comment and there could be  
853 several reasons for this. One explanation could be that we are simply under-estimating the  
854 uncertainty from the approximations made when calculating the Hessian for the observation  
855 operator. A more likely reason is non-linear effects in the treatment of uncertainties. In spite of our  
856 attempt to account for gross non-linear impacts through parameter transformations, residual non-  
857 linear effects may be causing this under-estimation of uncertainty by a factor of around 1.5.

858

859 FIGURE 3 ABOUT HERE

860 FIGURE 4 ABOUT HERE

861

### 862 **3.4 DA: Cloudy scenario**

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

873 lower quality dataset.

874

875 TABLE 9 ABOUT HERE

876

877 Table 9 shows the reduction in uncertainty for this experiment. One striking feature of these is that  
878 the percentage of cases within the credible interval is now above 80% in both cases, meaning that  
879 the reported uncertainties are close to the true values. Whilst the apparent reduction in uncertainty is  
880 apparently quite small (indeed, there is an increase in uncertainty for some state vector elements) at  
881 1.53 for the first order constraint and 1.14 for the second order, when weighed against the improved  
882 statistical representation, these rise to values directly comparable with the results from the previous  
883 experiment. The credible intervals shown in figure 3 are now realistic representations of the state  
884 vector elements and their uncertainties, achieved with only 50% of the samples of the previous  
885 experiment and with large data gaps, which is an important result.

886

887 Figure 4 shows the posterior correlation matrices (the inverse Hessian matrix) for the cloudy  
888 scenario. The general pattern of this matrix for the ‘complete’ scenario is rather similar so not  
889 shown here. Obviously, the correlation is unity along the leading diagonal. Another important  
890 feature is that the broad patterns of positive and negative correlations that we noted for the ‘single  
891 obs’ solutions remains here. There is negative correlation between  $s_1$  and all terms by TLAI. There  
892 is negative correlation between TLAI and  $TC_{ab}$  and  $TC_w$  but positive correlation with  $TC_{dm}$ . These  
893 patterns are consistent for both constraints used. We notice then that the application of the dynamic  
894 model (regularisation) in time does not remove the correlations arising from the inverse Hessian of  
895 the observation cost function, but rather it ‘spreads’ uncertainty correlation out in the time domain.  
896 This is particularly visible in the second order constraint matrix in figure 4 where we can clearly see  
897 this smoothing being greater where there are data gaps ( $s_1$  is a good example of that). Equally,  
898 where a part of the state vector has been strongly influenced by the regularisation (e.g. N for the

899 first order constraint) we see very high correlation at all time steps. Another interesting feature of  
900 this figure is the fact that for some state vector elements (e.g. N for the second order constraint) we  
901 can clearly see the influence of the periodic boundary condition).

902

## 903 **4 Discussion**

### 904 **4.1 The value of an EO-LDAS**

905 This paper outlines a scheme for a weak constraint data assimilation system, developed in the ESA  
906 EO-LDAS project, designed for integrating Earth Observation data from a variety of sources over  
907 arbitrary time scales, and through that to multiple spatial resolutions. It has the potential, via careful  
908 definition of the underlying model to be extended to spatial constraints, although this is not  
909 explored here. The scheme is designed to allow interface with process models, should they be  
910 available, though only an empirical regularisation model is shown in this paper. The core of the  
911 system is a set of constraints on: (i) prior estimates of the state vector; (ii) a linear model of the state  
912 vector; (iii) observation operator (RT model) predictions of a set of EO data and a DA scheme  
913 around these using an iterative bounded optimisation approach (L-BFGS-B).

914

915 In this paper, we have set up and run a synthetic data experiment with EO data mimicking those that  
916 might be provided by the MSI sensors on the forthcoming Sentinel-2 platforms. Experiments in DA  
917 are conducted for an idealised ‘full coverage’ scenario (5 day sampling) and for a ‘cloudy’ case  
918 (average around 10 day sampling but with large data gaps of up to 60 days). The results are  
919 compared to baseline experiments where we attempt to estimate the state variable trajectories over  
920 the course of a year for a subset of the total state variables (six elements per observation period).  
921 The prior term is used only very weakly here, although bounds are set to the state vector elements.  
922 Further, we assume that we have direct access to the surface reflectance (as opposed to top of  
923 atmosphere radiance), and that the noise on the observations is uncorrelated and of known  
924 magnitude. Broadly however, we can claim that the baseline results should be indicative of those

925 that might be obtained from Sentinel-2 data using ‘traditional’ estimation methods. For what we  
926 suppose to be a typical observation noise scenario, the uncertainty can be a quite large proportion of  
927 the signal for important terms such as LAI, this for a peak LAI of only around 3.7, although on  
928 average the uncertainty in TLAI may only be around 5%. This then, relates to the information  
929 content of a single MSI observation for this level of noise, assuming some important terms such as  
930 leaf angle distribution are known precisely. These results are not surprising but are simply a  
931 manifestation of the difficulty of the inference of biophysical parameters from remote radiometric  
932 observations: the problem may often be ill-posed (consider the situation if only two wavebands at  
933 red and near infrared were available), but even if it is not strictly so, there may not be sufficient  
934 information to very well constrain the information we require. In any case, there can be quite high  
935 correlation in uncertainty.

936

937 The ways to improve this situation are: (i) to obtain more observations (although more observations  
938 does not always translate to more information: consider again sampling at only red and near  
939 infrared wavelengths in trying to constrain e.g. leaf water content); or (ii) to add some other forms  
940 of information; (iii) average the data. Much *a priori* information has been used in the past to help  
941 constrain these problems, but this has often been approached in a rather *ad hoc* manner. Examples  
942 include: assuming some terms known, without considering the impact of uncertainty in these, or  
943 imposing degrees of smoothness; assuming that some terms are constant over some arbitrary time  
944 period; or *post hoc* low pass filtering to the final results. Given its success in other field of science  
945 and engineering, many authors have proposed that DA should be seen as the route to integration of  
946 the various forms of information one might wish to use to constrain the estimation. Key to DA is the  
947 weighting between the various sources of evidence, and key to this is assigning uncertainty  
948 correctly to the sources. This is a feature of the approach that dramatically differentiates it from the  
949 way in which VIs are mostly used in EO. As we note in the introduction, if we wish to estimate  
950 biophysical parameters (such as LAI) there is generally some form of calibration (against ground

951 observations or RT model runs) but it is extremely rare that those model uncertainties are  
952 considered in mapping the product. Other processing steps such as angular normalisation may have  
953 taken place, but again, any concepts of uncertainty arising from these are on the whole disregarded.  
954 All of these issues *could* be addressed within a DA framework, even if the source of the EO  
955 information were to be VIs.

956

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

975

976 The EO-LDAS scheme that we have built is capable of using any linearised process model and of

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

986  
987 We have also demonstrated (figure 2) that it is feasible to estimate the required hyper-parameters  
988 from some form of cross-validation exercise to impose an appropriate degree of model uncertainty,  
989 and that quite consistent results can be obtained even under cloudy conditions. This is an important  
990 practical point for the eventual operationalisation of these methods, but the area requires a little  
991 more discussion of practical issues in its implementation.

992  
993 Approximate linearization of the RT model variables here, following Weiss et al. (2000), has  
994 allowed Gaussian distributions to be assumed throughout. Although we have not directly  
995 investigated any residual non-linear effects in this study, some evidence is provided that on average  
996 we may be predicting only around 2/3 of the true uncertainty.

997  
998 **4.1 Future directions**

999 In this paper, we have only demonstrated DA for a homogeneous observation system, i.e. one for  
1000 which we have assumed the spectral sampling (and in effect, spatial resolution) for all observations  
1001 is the same. Using the EO-LDAS prototype for spectrally heterogeneous systems is straightforward,  
1002 but further work is needed to test the multi-scale concepts that would more generally be required.

1003 Within the existing prototype, the state vector can represent any mixture of temporal or spatial  
1004 samples. The concepts of temporal smoothness used here apply equally to the spatial domain  
1005 (indeed, such ideas form the basis of the field of geostatistics (e.g. Atkinson and Lewis, 2000)), so  
1006 the prototype can be used directly to link a state vector representation on a spatial grid, via  
1007 appropriate specification of the matrix  $A$ . Indeed, one could consider the experiments performed in  
1008 this paper simply as being on a spatial transect, rather than as we have assumed a temporal sampling  
1009 pattern. The only practical difference is that in that case, the viewing and illumination angles would  
1010 be near identical for all samples.

1011

1012 The EO-LDAS prototype is designed to allow a (relatively) large number of state variables to be  
1013 estimated simultaneously in a variational system ( $> 2000$  demonstrated here). One potential  
1014 advantage of this is that information can be passed between any of the state vector elements. In  
1015 practice, we have only used rather local information transfer in the model constraints applied here  
1016 (differences with neighbours in time) and this approach could also be implemented as a sequential  
1017 smoother. In viewing the temporal experiment we have performed as effectively equivalent to a  
1018 spatial experiment, the neighbourhood need not be very different (i.e., in the spatial sense, we could  
1019 follow the approach here and directly connect information in one grid cell to its 8 neighbours).  
1020 However, this variational system maintains the capacity for more distant (time or space)  
1021 connections, for example in applying multiple scale constraint.

1022

1023 A point that we have not dwelt on in this paper is the time required for processing. This is currently  
1024 around several hours for solving for  $> 2000$  state vector elements using 73 samples for what equates  
1025 to a single pixel (albeit for all samples over a year). The experiments in this paper were conducted  
1026 over around 120 UNIX cores, so quite large-scale experiments are feasible using University  
1027 computing resources. Clearly the processing requirements would need to be greatly reduced if such  
1028 a system were to be proposed for operational processing. The computer code is not on the whole



1029 written to be fast, but rather to be adequate to learn about using this form of DA. There are various  
1030 ways in which this might be tackled: clearly the very tight convergence criteria could be somewhat  
1031 relaxed, and more efficient codes could be written, but there will always be a relatively large  
1032 overhead on multiple calculations of a radiative transfer model. Pragmatic ways to overcome this  
1033 issue have mainly in the past dealt with using LUTs or ANNs to sample or approximate the  
1034 observation operator, but clearly in the DA framework we must consider representational error in  
1035 any such emulation. One avenue that holds much promise is that of Gaussian Process (GP)  
1036 emulators (Kennedy and O'Hagan, 2000, 2001), a form of regression that has been successfully  
1037 used to simulate computationally costly models runs through simple functional approximations. The  
1038 great benefit of this latter approach is that uncertainties in the emulated model are included and that  
1039 derivatives of the model can also be easily produced. If we consider the observation operator as a  
1040 sampled function with GP emulation, it is interesting to note that the underlying concepts implying  
1041 smooth interpolation with treatment of representation uncertainty are of course the same as we are  
1042 performing in the temporal (or indeed spatial) process model in the DA.

1043

## 1044 **5 Conclusions**

1045 The EO-LDAS prototype that is described in this paper has been demonstrated to be capable of  
1046 simultaneously estimating a state vector of over 2000 elements of surface biophysical  
1047 characteristics in a synthetic experiment using simulated Sentinel-2 MSI data. Although the  
1048 processing time required for this is currently a little long, this is a significant step in the size of such  
1049 problems that can be tackled simultaneously. The ability to do this derives from the use of AD-  
1050 generated adjoint code for the observation operator at the heart of the DA system.

1051

1052 The DA scheme that has been developed is a weak constraint variational system. The value of such  
1053 a scheme has been demonstrated using the synthetic MSI data to show a reduction in uncertainty of  
1054 up to around 2 when a linear dynamic model is used in the DA. The linear dynamic model is

1055 proposed as a general implementation that can potentially be interfaced to biophysical process  
1056 models through linearization. It is used in this paper with first and second-order derivative  
1057 constraints (zero- and first-order process models) which are shown to be sufficient to track rather  
1058 complex biophysical parameter trajectories via a radiative transfer model ‘observation operator’  
1059 interface to the synthetic EO data.

1060  
1061 We have noted at various points in this text, that some aspects of the EO-LDAS prototype are still  
1062 under development of testing, but what actually is provided by the prototype code is a functioning  
1063 tool for exploring many issues in DA and for estimating information on surface biophysical  
1064 parameters. The tool is designed as a weak constraint variational system, but we have argued that it  
1065 can also be used sequentially as it stands. We have demonstrated the use of the tool and of DA  
1066 concepts in reducing uncertainty in biophysical parameter estimation in a temporal sense, but also  
1067 argued the equivalence of this (in DA in general, but in the tool specifically) for the spatial domain  
1068 as well. We have used only empirical ‘regularisation’ concepts in demonstrating the DA, but noted  
1069 that these are powerful general concepts that are extremely useful, particularly if biophysical  
1070 models do not treat some of the parameters we are concerned with. In the more general case though,  
1071 any linearization of a more process-driven model can be directly interfaced to the EO-LDAS  
1072 prototype.

1073  
1074 There is clearly quite a long way to go from initial experiments with relatively slow computer codes  
1075 to an operational system for land data information extraction from EO, i.e. an operational EO-  
1076 LDAS, but the concepts explored here demonstrate the power and potential flexibility of such an  
1077 approach.

1078  
1079 **Acknowledgements**

1080 We gratefully acknowledge the support of ESA through the EO-LDAS project 22205/09/I-EC for

1081 funding this work. We also acknowledge the support of the (UK) National Environment Research  
1082 Council (NERC) National Centre for Earth Observation (NCEO) for its support of several of the  
1083 personnel involved in this work. We would further like to thank the attendees of the EO-LDAS  
1084 Community Workshop held in ESA ESRIN in November 2009 for their feedback and inputs to this  
1085 study.

1086

1087 **References**

1088 Asrar, G., Kanemasu, E.T., Jackson, R.D., Pinter Jr, P.J., 1985. Estimation of total above-ground  
1089 phytomass production using remotely sensed data. *Remote Sensing of Environment* 17, 211–220.

1090 Atkinson, P. and Lewis, P., 2000, *Geostatistical classification for remote sensing: an introduction*.  
1091 *Computers and Geoscience*, 26(4), 361-371.

1092 Baret, F. and Guyot, G., 1991. Potentials and limits of vegetation indices for LAI and APAR  
1093 assessment. *Remote Sensing of Environment* 35, 161–173.

1094 Behrenfeld, M.J., Randerson, J.T., McClain, C.R., Feldman, G.C., Los, S.O., Tucker, C.J.,  
1095 Falkowski, P.G., Field, C.B., Frouin, R., Esaias, W.E., others, 2001. Biospheric primary production  
1096 during an ENSO transition. *Science* 291, 2594.

1097 Beven, K., 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320, 18–36.

1098 Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained  
1099 optimization. *SIAM Journal on Scientific Computing* 16, 1190–1208.

1100 CEOS, 2011a. The CEOS Constellation for Land Surface Imaging [WWW Document]. URL  
1101 <http://wgiss.ceos.org/lcip/lpic.shtml>, accessed 18/11/2011.

1102 CEOS, 2011b, CEOS EO Handbook – Instrument summary: HRG. [WWW Document]. URL  
1103 <http://database.eohandbook.com/database/instrumentsummary.aspx?instrumentID=183> accessed  
1104 18/11/2011.

1105 Chen, J.M., Pavlic, G., Brown, L., Cihlar, J., Leblanc, S.G., White, H.P., Hall, R.J., Peddle, D.R.,  
1106 King, D.J., Trofymow, J.A., others, 2002. Derivation and validation of Canada-wide coarse-

1107 resolution leaf area index maps using high-resolution satellite imagery and ground measurements.  
1108 Remote Sensing of Environment 80, 165–184.

1109 Choudhury, B.J., 1987. Relationships between vegetation indices, radiation absorption, and net  
1110 photosynthesis evaluated by a sensitivity analysis. Remote Sensing of Environment 22, 209–233.

1111 Clerici, M., Vossbeck, M., Pinty, B., Kaminski, T., Taberner, M., Lavergne, T., Andredakis, I., 2010.  
1112 Consolidating the two-stream inversion package (JRC-TIP) to retrieve land surface parameters from  
1113 albedo products. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal  
1114 of 3, 286–295.

1115 Combal, B., Baret, F., Weiss, M., Trubuil, A., Mace, D., Pragnere, A., Myneni, R., Knyazikhin, Y.,  
1116 Wang, L., 2003. Retrieval of canopy biophysical variables from bidirectional reflectance:: Using  
1117 prior information to solve the ill-posed inverse problem. Remote Sensing of Environment 84, 1–15.

1118 Council of the European Union, 2010. Taking forward the European Space Policy.

1119 DMCII, 2010. DMC Constellation [WWW Document]. URL  
1120 [http://www.dmcii.com/about\\_us\\_constellation.htm](http://www.dmcii.com/about_us_constellation.htm), accessed 20/12/2010.

1121 Eilers, P.H.C., 2003. *A perfect smoother*, Analytical Chemistry, 75(14), pp. 3631-3636

1122 Enting, I.G., 2002. Inverse problems in atmospheric constituent transport. Cambridge University  
1123 Press.

1124 ESA, 2010. Mission Requirements Document GMES Sentinel-2 [WWW Document]. URL  
1125 [http://esamultimedia.esa.int/docs/GMES/Sentinel-2\\_MRD.pdf](http://esamultimedia.esa.int/docs/GMES/Sentinel-2_MRD.pdf), accessed 20/12/2010.

1126 European Commission, n.d. GEOSS: Policy relevance, Future Challenges, EU contribution to  
1127 GEOSS, Relevant documentation [WWW Document]. GEOSS. URL  
1128 [http://ec.europa.eu/research/environment/index\\_en.cfm?section=geo&pg=geoss](http://ec.europa.eu/research/environment/index_en.cfm?section=geo&pg=geoss), accessed  
1129 20/12/2010.

1130 Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical  
1131 implementation. Ocean dynamics 53, 343–367.

1132 Fang, H., Liang, S., Hoogenboom, G., Teasdale, J., Cavigelli, M., 2008a. Corn-yield estimation

1133 through assimilation of remotely sensed data into the CSM-CERES-Maize model. *International*  
1134 *Journal of Remote Sensing* 29, 3011.

1135 Fang, H., Liang, S., Townshend, J.R., Dickinson, R.E., 2008b. Spatially and temporally continuous  
1136 LAI data sets based on an integrated filtering method: Examples from North America. *Remote*  
1137 *Sensing of Environment* 112, 75-93.

1138 Féret, J.B., François, C., Asner, G.P., Gitelson, A.A., Martin, R.E., Bidet, L.P., Ustin, S.L., le Maire,  
1139 G., Jacquemoud, S., 2008. PROSPECT-4 and 5: Advances in the leaf optical properties model  
1140 separating photosynthetic pigments. *Remote Sensing of Environment* 112, 3030–3043.

1141 Fung, A.K. and Chen, K., 2010. *Microwave Scattering and Emission Models for Users*, 1st ed.  
1142 Artech House Publishers.

1143 Garrigues, S., Lacaze, R., Baret, F., Morisette, J.T., Weiss, M., Nickeson, J.E., Fernandes, R.,  
1144 Plummer, S., Shabanov, N.V., Myneni, R.B., others, 2008. Validation and intercomparison of global  
1145 Leaf Area Index products derived from remote sensing data. *Journal of Geophysical Research*, 113,  
1146 G02028.

1147 Ghil, M. and Malanotte-Rizzoli, P., 1991. Data assimilation in meteorology and oceanography. *Adv.*  
1148 *Geophys* 33, 141–266.

1149 Giering, R. and Kaminski, T., 1998. Recipes for adjoint code construction. *ACM Transactions on*  
1150 *Mathematical Software (TOMS)* 24, 437–474.

1151 Gill, P.E., Murray, W., Wright, M.H., 1981. *Practical optimization*. Academic Press, London and  
1152 New York.

1153 Gobron, N., Pinty, B., Verstraete, M.M., Govaerts, Y., 1997. A semidiscrete model for the scattering  
1154 of light by vegetation. *Journal of Geophysical Research* 102, 9431–9446.

1155 Gobron, N., Pinty, B., Verstraete, M., Widlowski, J., 2000. Advanced vegetation indices optimized  
1156 for up-coming sensors: Design, performance, and applications. *IEEE Transactions on Geoscience*  
1157 *and Remote Sensing*, 38, 2489-2505.

1158 Gobron, N., Pinty, B., Verstraete, M.M., Widlowski, J.L., 2002. Advanced vegetation indices

1159 optimized for up-coming sensors: Design, performance, and applications. *IEEE Transactions on*  
1160 *Geoscience and Remote Sensing*, 38, 2489–2505.

1161 Gobron, N., Belward, A., Pinty, B., Knorr, W., 2010. Monitoring biosphere vegetation 1998–2009.  
1162 *Geophysical Research Letters* 37, L15402.

1163 Goel, N.S., 1988. Models of vegetation canopy reflectance and their use in estimation of  
1164 biophysical parameters from reflectance data. *Remote Sensing Reviews* 4, 1–212.

1165 Goel, N.S., Thompson, R.L., 2000. A snapshot of canopy reflectance models and a universal model  
1166 for the radiation regime. *Remote Sensing Reviews* 18, 197–225.

1167 Goward, S.N., Tucker, C.J., Dye, D.G., 1985. North American vegetation patterns observed with the  
1168 NOAA-7 advanced very high resolution radiometer. *Plant Ecology* 64, 3–14.

1169 Jacquemoud, S., Baret, F., 1990. PROSPECT: A model of leaf optical properties spectra. *Remote*  
1170 *Sensing of Environment* 34, 75–91.

1171 Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. *Journal of the Royal*  
1172 *Statistical Society: Series B (Statistical Methodology)* 63, 425–464.

1173 Kennedy, M.C., O'Hagan, A., 2000. Predicting the output from a complex computer code when fast  
1174 approximations are available. *Biometrika* 87, 1.

1175 Kimes, D.S., Knyazikhin, Y., Privette, J.L., Abuelgasim, A.A., Gao, F., 2000. Inversion methods for  
1176 physically-based models. *Remote Sensing Reviews* 18, 381–439.

1177 Knorr, W., Kaminski, T., Scholze, M., Gobron, N., Pinty, B., Giering, R., Mathieu, P., 2010. Carbon  
1178 cycle data assimilation with a generic phenology model. *Journal of Geophysical. Research*, 115,  
1179 G04017.

1180 Kuusk, A., 1995. A fast, invertible canopy reflectance model. *Remote Sensing of Environment* 51,  
1181 342–350.

1182 Lavergne, T., Kaminski, T., Pinty, B., Taberner, M., Gobron, N., Verstraete, M.M., Vossbeck, M.,  
1183 Widlowski, J.L., Giering, R., 2007. Application to MISR land products of an RPV model inversion  
1184 package using adjoint and Hessian codes. *Remote Sensing of Environment* 107, 362–375.

1185 Lewis, P. and Disney, M., 2007. Spectral invariants and scattering across multiple scales from  
1186 within-leaf to canopy. *Remote Sensing of Environment* 109, 196–206.

1187 Liang, S. and Strahler, A.H., 2002. An analytic BRDF model of canopy radiative transfer and its  
1188 inversion. *IEEE Transactions on Geoscience and Remote Sensing*, 31, 1081–1092.

1189 Lu, X., Liu, R., Liu, J., Liang, S., 2007. Removal of noise by wavelet method to generate high  
1190 quality temporal data of terrestrial MODIS products. *Photogrammetric Engineering and Remote*  
1191 *Sensing* 73, 1129.

1192 Lubansky, A. S., Yeow, Y. L., Leong Y-K., Wickramasinghe S. R., Han B., 2006. A general method  
1193 of computing the derivative of experimental data., *AIChE J.*, 52, pp. 323-332.

1194 Lyapustin, A. and Knyazikhin, Y., 2001, Method of Green Function in the Radiative Transfer  
1195 Problem. Part I: Homogeneous non-Lambertian Surface. *Applied Optics*, 40, 3495-3501.

1196 Lyapustin, A., Wang, Y., Martonchik, J., Privette, J., Holben, B., Slutsker, I., Sinyuk, A. and  
1197 Smirnov, A., 2006, Local analysis of MISR surface BRDF and albedo over GSFC and Mongu  
1198 AERONET sites, *IEEE Transactions on Geoscience and Remote Sensing*, 44, 1707-1718.

1199 McLaughlin, D., 2002. An integrated approach to hydrologic data assimilation: interpolation,  
1200 smoothing, and filtering. *Advances in Water Resources* 25, 1275–1286.

1201 Myneni, R.B., Maggion, S., Iaquinta, J., Privette, J.L., Gobron, N., Pinty, B., Kimes, D.S.,  
1202 Verstraete, M.M., Williams, D.L., 1995. Optical remote sensing of vegetation: Modeling, caveats,  
1203 and algorithms. *Remote Sensing of Environment* 51, 169-188.

1204 NASA, 2010. NASA - A-Train [WWW Document]. URL  
1205 [http://www.nasa.gov/mission\\_pages/cloudsat/multimedia/a-train.html](http://www.nasa.gov/mission_pages/cloudsat/multimedia/a-train.html)

1206 Nemani, R.R., Keeling, C.D., Hashimoto, H., Jolly, W.M., Piper, S.C., Tucker, C.J., Myneni, R.B.,  
1207 Running, S.W., 2003. Climate-driven increases in global terrestrial net primary production from  
1208 1982 to 1999. *Science* 300, 1560.

1209 Olioso, A., Inoue, Y., Ortega-Farias, S., Demarty, J., Wigneron, J.P., Braud, I., Jacob, F.,  
1210 Lecharpentier, P., Ottlé, C., Calvet, J.C., others, 2005. Future directions for advanced

1211 evapotranspiration modeling: Assimilation of remote sensing data into crop simulation models and  
1212 SVAT models. *Irrigation and Drainage Systems* 19, 377–412.

1213 Pinty, B., Lavergne, T., Vossbeck, M., Kaminski, T., Aussedat, O., Giering, R., Gobron, N.,  
1214 Taberner, M., Verstraete, M.M. and Widlowski, J-L. (2007) Retrieving surface parameters for  
1215 climate models from Moderate Resolution Imaging Spectroradiometer (MODIS)-Multiangle  
1216 Imaging Spectroradiometer (MISR) albedo products, *Journal of Geophysical Research*, 112,  
1217 D10116, doi:10.1029/2006JD008105.

1218 Price, J.C., 1990. On the information content of soil reflectance spectra. *Remote sensing of*  
1219 *Environment* 33, 113–121.

1220 Privette, J.L., Myneni, R.B., Tucker, C.J., Emery, W.J., 1994. Invertibility of a 1-D discrete  
1221 ordinates canopy reflectance model. *Remote Sensing of Environment* 48, 89–105.

1222 Qin, J., Liang, S., Li, X., Wang, J., 2008. Development of the adjoint model of a canopy radiative  
1223 transfer model for sensitivity study and inversion of leaf area index. *IEEE Transactions on*  
1224 *Geoscience and Remote Sensing*, 46, 2028–2037.

1225 Qin, J., Liang, S., Liu, R., Zhang, H., Hu, B., 2007. A weak-constraint-based data assimilation  
1226 scheme for estimating surface turbulent fluxes. *IEEE Geoscience and Remote Sensing Letters*, 4,  
1227 649–653.

1228 Qin, J., Liang, S., Yang, K., Kaihotsu, I., Liu, R., Koike, T., 2009. Simultaneous estimation of both  
1229 soil moisture and model parameters using particle filtering method through the assimilation of  
1230 microwave signal. *Journal of Geophysical Research* 114, D15103.

1231 Quaife, T., Lewis, P., 2010. Temporal Constraints on Linear BRDF Model Parameters. *IEEE*  
1232 *Transactions on Geoscience and Remote Sensing*, 48, 2445–2450.

1233 Quaife, T., Lewis, P., De Kauwe, M., Williams, M., Law, B.E., Disney, M., Bowyer, P., 2008.  
1234 Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter.  
1235 *Remote Sensing of Environment* 112, 1347-1364.

1236 Richardson, A.J., Wiegand, C.L., 1977. Distinguishing vegetation from soil background



1237 information(by gray mapping of Landsat MSS data). *Photogrammetric Engineering and Remote*  
1238 *Sensing* 43, 1541–1552.

1239 Rochdi, N., Fernandes, R., 2010. Systematic mapping of Leaf Area Index across Canada using 250-  
1240 meter MODIS data. *Remote Sensing of Environment* 114, 1130–1135.

1241 Rodgers, C.D., 2000. *Inverse Methods for Atmospheric Sounding : Theory and Practice*. World  
1242 Scientific Publishing Company.

1243 Roy, D.P., Jin, Y., Lewis, P.E., Justice, C.O., 2005. Prototyping a global algorithm for systematic  
1244 fire-affected area mapping using MODIS time series data. *Remote Sensing of Environment* 97, 137-  
1245 162.

1246 Slater, A.G. and Clark, M.P., 2009. Snow data assimilation via an ensemble Kalman filter. *Journal of*  
1247 *Hydrometeorology*, 7, 478-492.

1248 Sobrino, J.A., Jiménez-Muñoz, J.C., Verhoef, W., 2005. Canopy directional emissivity: Comparison  
1249 between models. *Remote Sensing of Environment* 99, 304–314.

1250 Stöckli, R., Rutishauser, T., Dragoni, D., O'Keefe, J., Thornton, P.E., Jolly, M., Lu, L., Denning,  
1251 A.S., 2008. Remote sensing data assimilation for a prognostic phenology model. *Journal of*  
1252 *Geophysical Research*, 113, 19 PP.

1253 Tarantola, A., 2005. *Inverse problem theory and methods for model parameter estimation*. Society  
1254 for Industrial and Applied Mathematics.

1255 Tha Paw, U., others, 1992. Development of models for thermal infrared radiation above and within  
1256 plant canopies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 47, 189–203.

1257 Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation.  
1258 *Remote sensing of Environment*, 8, 127–150.

1259 Twomey, S., 2002. *Introduction to the Mathematics of Inversion in Remote Sensing*. Courier Dover  
1260 Publications.

1261 Verhoef, W., Bach, H., 2003. Simulation of hyperspectral and directional radiance images using  
1262 coupled biophysical and atmospheric radiative transfer models. *Remote Sensing of Environment*,

1263 87, 23–41.

1264 Vermote, E.F., Tanré, D., Deuzé, J.L., Herman, M., Morcrette, J.J., 1997. Second Simulation of the  
1265 Satellite Signal in the Solar Spectrum, 6S: An Overview. *IEEE Transactions on Geoscience and*  
1266 *Remote Sensing*, 35(3), 675-686.

1267 Vermote, E.F., El Saleous, N.Z., Justice, C.O., 2002. Atmospheric correction of MODIS data in the  
1268 visible to middle infrared: first results. *Remote Sensing of Environment*, 83, 97–111.

1269 Wahba, G., 1990. Spline models for observational data. Philadelphia, Pa: Society for Industrial and  
1270 Applied Mathematics.

1271 Weiss, M., Baret, F., Myneni, R.B., Pragnère, A., Knyazikhin, Y., 2000. Investigation of a model  
1272 inversion technique to estimate canopy biophysical variables from spectral and directional  
1273 reflectance data, *Agronomie* 20, 3–22.

1274 Widlowski, J.L., Taberner, M., Pinty, B., Bruniquel-Pinel, V., Disney, M., Fernandes, R., Gastellu-  
1275 Etchegorry, J.P., Gobron, N., Kuusk, A., Lavergne, T., others, 2007. Third Radiation Transfer Model  
1276 Intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models. *Journal of*  
1277 *Geophysical Research*, 112, D09111.

1278 Xiao, Z., Liang, S., Wang, J., Song, J., Wu, X., 2009. A temporally integrated inversion method for  
1279 estimating leaf area index from MODIS data. *IEEE Transactions on Geoscience and Remote*  
1280 *Sensing*, 47, 2536–2545.

1281 Xiao, Z., Liang, S., Wang, J., Jiang, B., Li, X., 2011. Real-time retrieval of Leaf Area Index from  
1282 MODIS time series data. *Remote Sensing of Environment* 115, 97-106.

1283 Zhu, C., Byrd, R.H., Lu, P., Nocedal, J., 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for  
1284 large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*  
1285 23, 550–560.

1286 Zupanski, D., 1997, A general weak constraint applicable to operational 4dvar data assimilation  
1287 systems, *Monthly weather review*, 125, 2274–2292.

1288

1289

1290

1291 **Table Captions**

1292 Table 1. Spatial resolution, Central wavelength and bandwidths for Sentinel-2 MSI (ESA, 2010).

1293 Table 2. Summary of observation operator state variables.

1294 Table 3. Transformations applied to approximate linearise state variable response.

1295 Table 4. Upper and lower bounds for the state vector terms (in transformed space, where  
1296 appropriate) used in the simulations, along with the temporal trajectory assumed.

1297 Table 5. Model uncertainty  $\gamma$  for each parameter, calculated from the synthetic model state vector.

1298  $TC_{dm}$  and  $N$  were kept constant, so have no theoretical model uncertainty associated.

1299 Table 6: Mean posterior uncertainty. Figures refer to the complete daily time series, while figures in  
1300 brackets refer to the mean posterior uncertainty only considering the dates where observations are  
1301 available.

1302 Table 7: Single observation posterior correlation matrix. Elements above the main diagonal show  
1303 the results for DoY 186, whereas the elements below the main diagonal represent the median of all  
1304 dates.

1305 Table 8: Uncertainty reduction relative to the single observation inversion, as well as percentage of  
1306 cases where the true parameter lies within the estimated 95% credible interval. Results for non-  
1307 cloudy scenario are reported. both complete time series.

1308 Table 9: Uncertainty reduction relative to the single observation, as well as percentage of cases  
1309 where the true parameter lies within the estimated 95% confidence interval. Results for cloudy  
1310 scenario.

1311

1312

1313

1314

1315

1316 .

1317

1318 **Figure Captions**

1319 Figure 1. Base level state vector estimated from inverting single observations, (left column) and for  
1320 model uncertainty unknown and estimated through cross-validation – first difference constraint  
1321 (central column) and second difference constraint (third column). Results for each of the six  
1322 parameters are shown in rows. True values are shown as a dashed line. The full lines are the  
1323 posterior means, and the shaded area represents the associated  $\pm 1.96$  standard deviations interval.  
1324 MSI observations are shown as open symbols. Crosses along the bottom of the third row indicate  
1325 the location of the cross validation acquisition dates.

1326 Figure 2. Error in cross validation scaled by observational uncertainty for varying model uncertainty  
1327  $\gamma$  for first and second order constraints. Vertical lines around 200 represent the theoretical value of  $\gamma$   
1328 for each of the 4 time-varying state variables using a first order constraint, and vertical lines around  
1329 5000 represent the theoretical values for  $\gamma$  for each of the 4 time-varying state variables using  
1330 second order constraint.

1331 Figure 3. Base level state vector estimated from inverting single observations, (left column) and for  
1332 model uncertainty unknown and estimated through cross-validation – first difference constraint  
1333 (central column) and second difference constraint (third column). Reduced number of acquisitions  
1334 due to cloud cover scenario. Results for each of the six parameters are shown in rows. True values  
1335 are shown as a dashed line. The full lines are the posterior means, and the shaded area represents the  
1336 associated  $\pm 1.96$  standard deviations interval. MSI observations are shown as open symbols.  
1337 Crosses along the bottom of the third row indicate the location of the cross validation acquisition  
1338 dates.

1339 Figure 4. Posterior correlation matrices for the cloudy scenario. Labels indicate the location of the  
1340 first day for component of the state vector.

1341

1342

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

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

1343

1344

1345

*Table 2. Summary of observation operator state variables.*

1346

1347

#	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 <sub>ab</sub>	μgcm <sup>-2</sup>	40	0	200
5	Carotenoids	C <sub>ar</sub>	μgcm <sup>-2</sup>	0	0	200
6	Leaf water	C <sub>w</sub>	cm <sup>-1</sup>	0.01	0.00001	0.04
7	Dry matter	C <sub>m</sub>	gcm <sup>-2</sup>	0.01	0.00001	0.02
8	Leaf layers	N	none	1.0	1.0	2.5
9	soil PC 1	s <sub>1</sub>	none	0.2	0.05	0.4
10	soil PC 2	s <sub>1</sub>	none	0	-0.1	0.1
11	soil PC 3	s <sub>1</sub>	none	0	-0.05	0.05
12	soil PC 4	s <sub>1</sub>	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

1348  
1349  
1350  
1351

*Table 3. Transformations applied to approximate linearise state variable response*

#	Transformed Symbol	Transformation
1	TLAI	$\exp(-LAI/2.0)$
4	TC <sub>ab</sub>	$\exp(-C_{ab}/100)$
5	TC <sub>ar</sub>	$\exp(-C_{ar}/100)$
6	TC <sub>w</sub>	$\exp(-50 C_w)$
7	TC <sub>dm</sub>	$\exp(-100 C_{dm})$

1352

1353



1354

*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
<b>1</b>	TLAI	0.067	0.995	$LAI = 0.21 + 3.51 \sin^5(\pi t)$
<b>4</b>	$TC_{ab}$	0.135	1.0	$C_{ab} = 10.5 + 208.7 t : t \leq 0.5$ $C_{ab} = 219.2 - 208.7 t : t \geq 0.5$
<b>6</b>	$TC_w$	0.135	1.0	$C_w = 0.068 +$ $0.020(\sin(\pi t + 0.1) * \sin(6\pi t + 0.1))$
<b>7</b>	$TC_{dm}$	0.135	1.0	$C_{dm} = 0.01$
<b>8</b>	N	1	2.5	$N = 1$
<b>9</b>	$s_l$	0.001	0.4	$s_l = 0.20 + 0.18(\sin(\pi t) * \sin(6\pi t))$

1355

1356

1357

1358

1359

1360

1361

*Table 5. Model uncertainty  $\gamma$  for each parameter, calculated from the synthetic model state vector.  $TC_{dm}$  and  $N$  were kept constant, so have no theoretical model uncertainty associated.*

<b>#</b>	<b>Symbol</b>	<b>First difference uncertainty</b>	<b>Second difference uncertainty</b>
<b>1</b>	TLAI	188	8298
<b>4</b>	$TC_{ab}$	303	7315
<b>6</b>	$TC_w$	132	2277
<b>9</b>	$s_l$	212	3861

1362

1363

1364

*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.*

1365

Symbol	Non-cloudy			Cloudy		
	Uncertainty Single Obs.	Uncertainty 1 <sup>st</sup> Diff	Uncertainty 2 <sup>nd</sup> Diff	Uncertainty Single Obs.	Uncertainty 1 <sup>st</sup> Diff	Uncertainty 2 <sup>nd</sup> 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 <sub>ab</sub>	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 <sub>w</sub>	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 <sub>dm</sub>	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)
$s_I$	0.17 (0.04)	0.02 (0.02)	0.03 (0.03)	0.20 (0.04)	0.04 (0.03)	0.05 (0.03)

1366

1367

1368

1369

1370

1371

1372

1373

*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 <sub>ab</sub>	TC <sub>w</sub>	TC <sub>dm</sub>	$N$	$s_I$
TLAI	1.00	0.16	-0.05	0.47	-0.25	0.58
TC <sub>ab</sub>	-0.44	1.00	0.15	-0.11	-0.47	0.34
TC <sub>w</sub>	-0.42	0.35	1.00	0.04	0.01	-0.14
TC <sub>dm</sub>	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
$s_I$	0.76	-0.53	-0.40	-0.25	-0.28	1.00

1374

*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 are reported. both complete time series*

#	Symbol	Complete time series				Observations only				
		Unc. red 1 <sup>st</sup> diff	Unc. red. 2 <sup>nd</sup> diff	% cases (1 <sup>st</sup> diff)	% cases (2 <sup>nd</sup> diff)	Unc. red 1 <sup>st</sup> diff	Unc. red. 2 <sup>nd</sup> diff	% cases (1 <sup>st</sup> diff)	% cases (2 <sup>nd</sup> 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 <sub>ab</sub>	5.24	3.58	61.1	65.2	2.57	1.74	60.3	65.8	65.8
6	TC <sub>w</sub>	3.47	1.77	51.2	69.9	2.72	1.38	50.7	71.2	60.3
7	TC <sub>dm</sub>	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	$s_I$	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

1376

1377

1378

*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.*

#	Symbol	Complete time series				Observations only				
		Unc. red 1 <sup>st</sup> diff	Unc. red. 2 <sup>nd</sup> diff	% cases (1 <sup>st</sup> diff)	% cases (2 <sup>nd</sup> diff)	Unc. red 1 <sup>st</sup> diff	Unc. red. 2 <sup>nd</sup> diff	% cases (1 <sup>st</sup> diff)	% cases (2 <sup>nd</sup> diff)	% cases (single)
1	TLAI	3.33	2.39	82.7	74	0.965	0.68	88.9	91.7	63.9
4	TC <sub>ab</sub>	3.68	2.93	80.0	89.6	1.82	1.58	61.1	83.3	58.3
6	TC <sub>w</sub>	2.33	1.4	64.1	85.2	1.90	1.18	83.3	88.9	61.1
7	TC <sub>dm</sub>	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	$s_I$	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

1379

1380

1381

1382

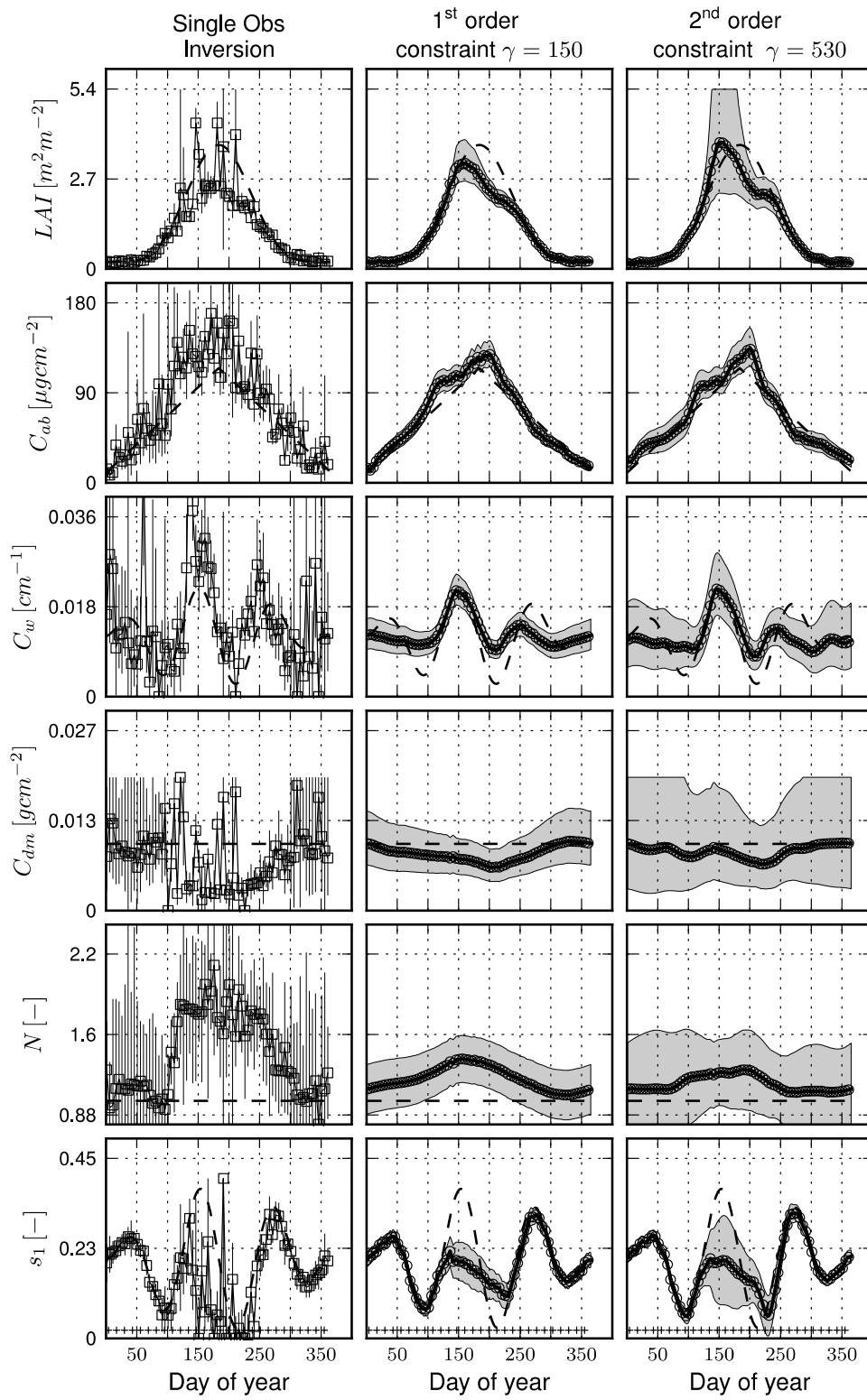
1383

*Figures*

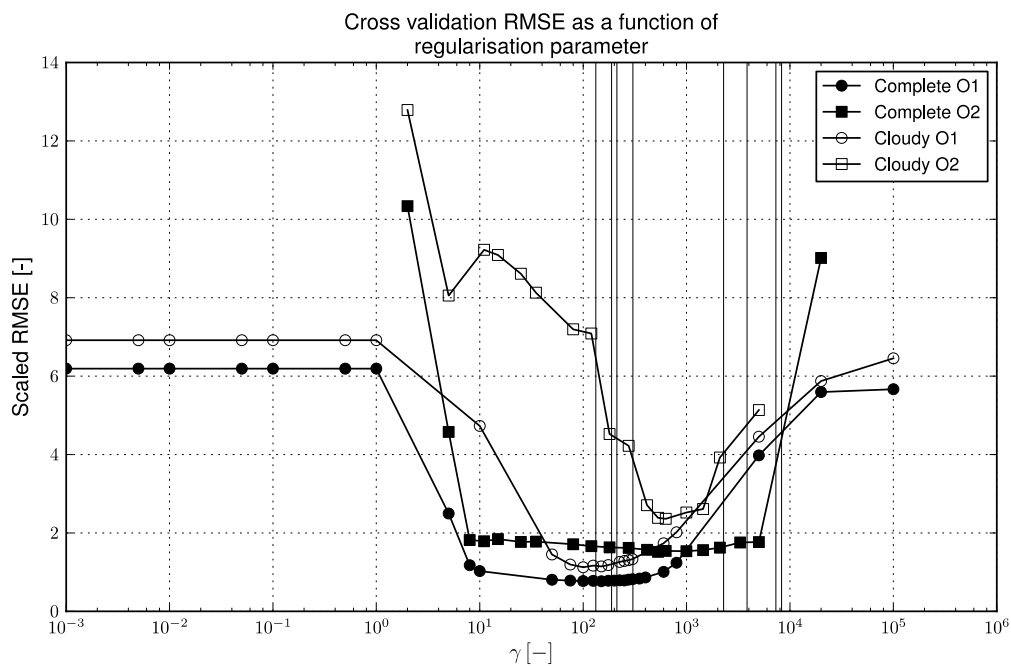
*Figure 1. Base level state vector estimated from inverting single observations, (left column) and for model uncertainty unknown and estimated through cross-validation – first difference constraint (central column) and second difference constraint (third column). Results for each of the six parameters 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  $\pm 1.96$  standard deviations interval. MSI observations are shown as open symbols. Crosses along the bottom of the third row indicate the location of the cross validation acquisition dates.*

1384

1385



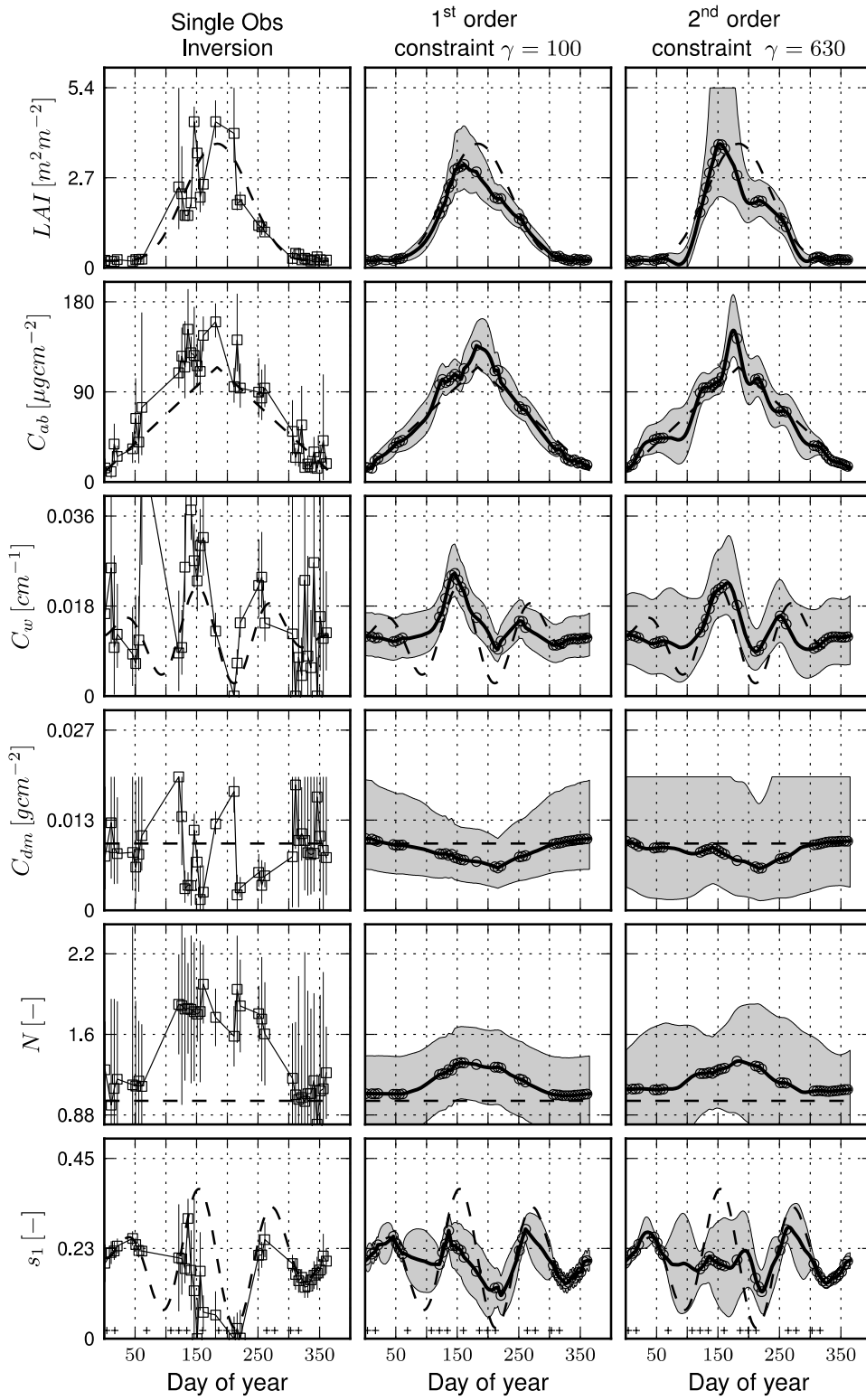
1387 **Figure 2. Error in cross validation scaled by observational uncertainty for varying model**  
 1388 **uncertainty  $\gamma$  for first and second order constraints. Vertical lines around 200 represent the**  
 1389 **theoretical value of  $\gamma$  for each of the 4 time-varying state variables using a first order**  
 1390 **constraint, and vertical lines around 5000 represent the theoretical values for  $\gamma$  for each of the**  
 1391 **4 time-varying state variables using second order constraint.**



1392



*Figure 3. 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  $\pm 1.96$  standard deviations interval. MSI observations are shown as open symbols. Crosses along the bottom of the third row indicate the location of the cross validation acquisition dates.*



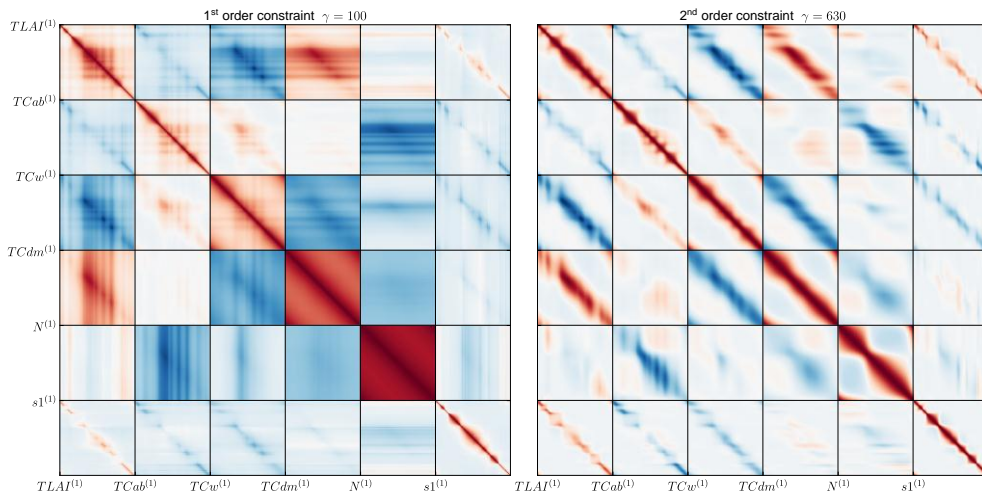
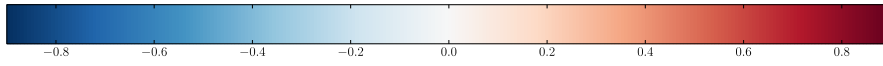
1393

1394

1395

1396

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



1397  
1398