Geoscientific
Model Development

# Constraining DALECv2 using multiple data streams and ecological constraints: analysis and application

Sylvain Delahaies[1], Ian Roulstone[1], and Nancy Nichols[2]

[1]Department of Mathematics, University of Surrey, Guildford, UK
[2]Department of Mathematics, University of Reading, Reading, UK

*Correspondence to:* Sylvain Delahaies (s.b.delahaies@surrey.ac.uk)

**Abstract.** We use a variational method to assimilate multiple data streams into the terrestrial ecosystem carbon cycle model DALECv2 (Data Assimilation Linked Ecosystem Carbon). Ecological and dynamical constraints have recently been introduced to constrain unresolved components of this otherwise ill-posed problem. Here we recast these constraints as a multivariate Gaussian distribution to incorporate them into the variational framework and we demonstrate their advantage through a linear analysis. Using an adjoint method we study a linear approximation of the inverse problem: firstly we perform a sensitivity analysis of the different outputs under consideration, and secondly we use the concept of resolution matrices to diagnose the nature of the ill-posedness and evaluate regularisation strategies. We then study the non-linear problem with an application to real data. Finally, we propose a modification to the model: introducing a spin-up period provides us with a built-in formulation of some ecological constraints which facilitates the variational approach.

## 1 Introduction

Carbon is a fundamental constituent of life and understanding its global cycle is a key challenge for the modelling of the Earth system. Through the processes of photosynthesis and respiration, ecosystems play a major role in the carbon cycle and thus in the dynamics of the global climate system. Our knowledge of the biogeochemical processes of ecosystems and an ever-growing amount of Earth observation systems can be combined using inverse modelling strategies to improve model predictions and uncertainty quantification.

The Data Assimilation Linked Ecosystem Carbon (DALEC) model is a simple box model for terrestrial ecosystems simulating a large range of processes occurring at different timescales from days to millennia. The work of Williams et al. (2005) established the benefit of using DALEC together with net ecosystem exchange (NEE) of $CO_2$ measurements in a Bayesian framework to estimate initial carbon stocks and model parameters, to improve flux predictions for ecosystem models and to quantify uncertainties. Inter-comparison experiments (Fox et al., 2009; Hill et al., 2012) have then demonstrated the relative merit of various inverse modelling strategies using NEE and MODIS leaf area index observations: most results agreed on the fact that parameters and initial stocks directly related to fast processes were best estimated with narrow confidence intervals, whereas those related to slow processes were poorly estimated with very large uncertainties. Other studies have tried to overcome this difficulty by adding complementary data streams (see Richardson et al. (2010)) or by considering longer observation windows (see Hill et al. (2012)). Recently Bloom and Williams (2015) defined a set of ecological and dynamical constraints (EDCs) to reject unrealistic parameter combinations in the absence of additional data. However, to date very few systematic analysis has been carried out to explain the large differences among results.

As with many inverse problems, assimilating Earth observations into DALEC is an ill-posed problem: the model–observation operator which relates parameters and initial carbon stocks to the observations is rank deficient and not all variables can be estimated, or the model–observation operator is ill-conditioned and small observational noise may lead

to a solution we can have little confidence in. Solving the problem amounts first to transforming it into a tractable problem in order to ensure a robust, meaningful and stable solution. This can be achieved by using regularisation techniques; the most popular one involves combining the observations and prior information, assuming it exists, through Bayesian inference. The choice of regularisation method depends on the nature of the problem and on the inverse modelling approach adopted.

So far, off-the-shelf methods such as ensemble Kalman filter (EnKF) and Monte Carlo Markov Chain (MCMC) were adopted to perform model–data fusion with DALEC. For its ability to accommodate non-linearity and any kind of probability distributions, the MCMC method, in the limit of a large number of samples, may be considered as the gold standard. However, despite being well suited for this type of small-scale problem, the computational complexity of MCMC method makes it intractable for more complex situations. Here we adopt a variational approach (4DVAR) where a cost function measuring the mismatch between the model and observations is minimised using a gradient method based on the adjoint of the model. At AmeriFlux sites (see http://ameriflux.lbl.gov/), we use MODIS monthly mean leaf area index (LAI) observations over a 12-year time window together with flux tower measurements of NEE and gross primary production (GPP). 4DVAR facilitates the diagnosis of the ill-posedness of the inverse problem: using model resolution matrices we can assess the resolution and stability properties of the observation operators and of the regularisation terms. We transcribe the EDCs into a novel variational framework and use some of this additional knowledge to estimate the otherwise undetermined variables. We consider a modification of the DALEC model by adding a spin-up period where carbon stocks are brought to equilibrium. This offers an alternative to including all the EDCs and helps reducing the confidence intervals for the predicted fluxes.

The paper is organised as follows. In Sect. 2 we present DALECv2 and the observation streams used in this study, review the EDCs introduced in Bloom and Williams (2015) and perform a sensitivity analysis of the different outputs of DALECv2 of interest for our experiments. In Sect. 3 we recall basic principles of inverse theory from a Bayesian perspective, we introduce the variational formulation and we show how to incorporate the EDCs into this framework. Section 4 is devoted to a résumé of the linearised problem, using the tangent linear model, where the challenges of ill-posed problems and their regularisation can be explored in detail using simple linear algebra. Using a singular value decomposition we illustrate the effect of observational noise on ill-conditioned systems, and we investigate solution strategies from the point of view of resolution matrices. In Sect. 5 we conduct a series of non-linear inverse modelling experiments using multiple data streams and EDCs. In Sect. 6 we modify DALECv2 to include a spin-up period which offers a built-in formulation of some EDCs, and then we reproduce the non-
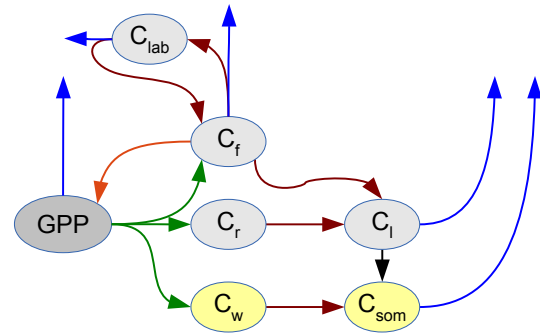


**Figure 1.** DALECv2 links the carbon pools (C) via allocation fluxes (green), litterfall fluxes (red) and decomposition (black). Respiration is represented by the blue arrows. The orange arrow represents the feedback of foliar carbon to gross primary production (GPP).

linear experiments. In Sect. 7 we discuss several extension to our manuscript and finally in Sect. 8 we draw conclusions.

## 2   Model, constraints and observations

### 2.1   DALECv2

DALECv2 depicts a terrestrial ecosystem as a set of six carbon pools (labile $C_{lab}$, foliar $C_f$, wood $C_w$, root $C_r$, litterfall $C_l$ and soil organic matter $C_s$) linked via allocation fluxes. At a monthly time step the gross primary production (GPP) is calculated using the Aggregated Canopy Model (Williams et al., 1997) as a non-linear function of meteorological drivers (temperature, radiation, atmospheric $CO_2$ concentration), foliar carbon and foliar nitrogen. Following the mass conservation principle, GPP is then allocated to the different carbon pools or released in the atmosphere via respiration. The schematic for DALECv2 is represented in Fig. 1 and a complete description of the model can be found in Bloom and Williams (2015). DALECv2 combines the two previous DALEC-evergreen and DALEC-deciduous models into a single model where the non-differentiable phenology process of DALEC-deciduous has been replaced with a differentiable process. DALECv2 is a non-linear dynamical system and the carbon pools are dynamical variables parametrised by their initial values $C_0$ and by 17 parameters $p$, whose range and description can be found in Table 1. The magnitudes and ranges of the parameters and the initial values vary drastically; therefore, to avoid the computational problems caused by these different scales the variational methods will be formulated and implemented in terms of the log transformed variable $x = \log([p, C_0])^T$. However, in order to limit unnecessary notation and definition, in the remainder of this paper $p$ and $C_0$ will stand for their log transform.

**Table 1.** DALECv2 dynamical variables and parameters with their respective range. The units of the non-dimensionless quantities are given in brackets.

| Label | Variable | Description | Range |
|---|---|---|---|
| $C_0(1)$ | $C_{lab}$ | initial labile C pool ($gC\,m^{-2}$) | 20–2000 |
| $C_0(2)$ | $C_f$ | initial foliar C pool ($gC\,m^{-2}$) | 20–2000 |
| $C_0(3)$ | $C_r$ | initial fine root C pool ($gC\,m^{-2}$) | 20–2000 |
| $C_0(4)$ | $C_w$ | initial above and below ground woody C pool ($gC\,m^{-2}$) | $100–10^5$ |
| $C_0(5)$ | $C_l$ | initial litter C pool ($gC\,m^{-2}$) | 20–2000 |
| $C_0(6)$ | $C_s$ | initial soil organic matter C pool ($gC\,m^{-2}$) | $100–2 \times 10^5$ |
| $p_1$ | $\theta_{min}$ | litter mineralisation rate ($day^{-1}$) | $10^{-5}–10^{-2}$ |
| $p_2$ | $f_a$ | autotrophic respiration fraction | 0.3–0.7 |
| $p_3$ | $f_f$ | fraction of GPP allocated to $C_f$ | 0.01–0.5 |
| $p_4$ | $f_r$ | fraction of GPP allocated to $C_r$ | 0.01–0.5 |
| $p_5$ | $c_{lf}$ | annual leaf loss fraction (season) | 1 - 8 |
| $p_6$ | $\theta_w$ | $C_w$ turnover rate ($day^{-1}$) | $2.5 \times 10^{-5}–10^{-3}$ |
| $p_7$ | $\theta_r$ | $C_r$ turnover rate ($day^{-1}$) | $10^{-4}–10^{-2}$ |
| $p_8$ | $\theta_l$ | $C_l$ turnover rate ($day^{-1}$) | $10^{-4}–10^{-2}$ |
| $p_9$ | $\theta_s$ | $C_s$ turnover rate ($day^{-1}$) | $10^{-7}–10^{-3}$ |
| $p_{10}$ | $\Theta$ | temperature dependence exponent factor | 0.018–0.08 |
| $p_{11}$ | $c_{eff}$ | canopy efficiency parameter | 10 - 100 |
| $p_{12}$ | $d_{onset}$ | leaf onset day (day) | 1–365 |
| $p_{13}$ | $f_l$ | fraction of GPP allocated to $C_{lab}$ | 0.01–0.5 |
| $p_{14}$ | $c_{ronset}$ | $C_{lab}$ release period (days) | 10–100 |
| $p_{15}$ | $d_{fall}$ | leaf fall day (day) | 1–365 |
| $p_{16}$ | $c_{rfall}$ | leaf fall period (days) | 20–150 |
| $p_{17}$ | $c_{lma}$ | leaf mass per area ($gC\,m^{-2}$) | 10–400 |

The meteorological drivers are extracted from $0.125° \times 0.125°$ ERA-Interim reanalysis data sets. For the purpose of our inverse modelling experiments we use four different observation streams: LAI, NEE, GPP and RESP (total respiration). LAI monthly mean observations for AmeriFlux sites are extracted from MOD15A2 LAI 8-day version 005 1 km resolution product. These observations together with the meteorological drivers are provided by A. Bloom and J. Exbrayat. Details about their construction can be found in Bloom and Williams (2015). At AmeriFlux sites we use the level 4 data product (available at http://cdiac.ornl.gov/ftp/ameriflux/data/Level4/), which provides monthly means for NEE and GPP. NEE and GPP are then used to define RESP as RESP = NEE + GPP. The meteorological drivers span a period of 12 years from 2001 to 2013. LAI observations are available during the full period but for NEE and GPP, and thus RESP, shorter records are available depending on the AmeriFlux site. In this study we consider the Morgan Monroe State Forest located in Indiana, USA (39.3–86.4). This AmeriFlux site is composed in majority of mixed hardwood broadleaf deciduous trees and classifies as a humid subtropical climate.

In the remainder of the paper the main focus is on the vector $x = \log([p, C_0])^T$: in Sect. 2.3 first where we investigate the sensitivity of different outputs with respect to $x$ and its components, and then in subsequent sections where $x$ is es-timated using inverse methods. The vector $x$, denoting fixed quantities as initial conditions and parameters for the dynamical system DALECv2, is seen as the variable from the point of view of sensitivity analysis and inverse modelling and therefore its components will be referred to as state variables, input variables or parameters interchangeably throughout the manuscript.

## 2.2 Ecological constraints

Over the last decade many inverse modelling studies have used NEE measurements from the FLUXNET network, together with other types of observations when available, to provide information about processes controlled by parameters with respect to which NEE is weakly sensitive. Though it contains an ever-increasing amount of information, the flux tower network only provides sparse coverage of terrestrial ecosystems. On the other hand, despite good spatial and temporal coverage, MODIS LAI monthly mean observations only constrain a limited set of DALECv2 state variables, and additional information is required in order to regularise the ill-posed problem and obtain a meaningful solution.

Additional information can be obtained by imposing priors on the variables or by adding other observation streams (biomass, soil organic matter, etc.). As an alternative, Bloom and Williams (2015) introduced a set of constraints, referred to as ecological and dynamical constraints (EDCs). These

constraints, detailed in Bloom and Williams (2015), can be divided into two groups: static and dynamic constraints. The static constraints which directly impose conditions on the parameters are as follows:

– Turnover rate constraints which ensure that turnover rates ratios are consistent with knowledge of the carbon pools residence times.

$$\text{EDC}_1: \quad p_9 < p_8, \tag{1}$$
$$\text{EDC}_2: \quad p_9 < p_1, \tag{2}$$
$$\text{EDC}_3: \quad p_6 < 1/(p_5 \times 365.25), \tag{3}$$
$$\text{EDC}_4: \quad p_7 > p_9 \exp p_{10}\overline{T}, \tag{4}$$
$$\text{EDC}_5: \quad p_{12} + 45 < p_{15}, \tag{5}$$

where $\overline{T}$ denotes the mean temperature within the drivers time window. $\text{EDC}_4$ is a modification to the constraint proposed in Bloom and Williams (2015). It is currently used in the CARDAMON framework (http://www.geos.ed.ac.uk/homes/mwilliam/CARDAMOM.html).

– Root–foliar allocation which allows for a strong correlation between parameters controlling allocation to foliage and roots.

$$\text{EDC}_6: \quad f_{\text{root}} < 5\,(f_{\text{fol}} + f_{\text{lab}}), \tag{6}$$
$$\text{EDC}_7: \quad f_{\text{fol}} + f_{\text{lab}} < 5 f_{\text{root}}, \tag{7}$$

where the allocation fractions $f_{\text{fol}}$, $f_{\text{lab}}$ and $f_{\text{root}}$ are defined as

$$f_{\text{auto}} = p_2, \tag{8}$$
$$f_{\text{fol}} = (1 - f_{\text{auto}})\, p_3, \tag{9}$$
$$f_{\text{lab}} = (1 - f_{\text{auto}} - f_{\text{fol}})\, p_{13}, \tag{10}$$
$$f_{\text{root}} = (1 - f_{\text{auto}} - f_{\text{fol}} - f_{\text{lab}})\, p_4. \tag{11}$$

The dynamic constraints, for which a model run is performed to define attractors, limit the application of the model to ecosystems with no major recent disturbance. They are defined as follows:

– Root–foliar mean dynamics

$$\text{EDC}_8: \overline{C}_r < 5\overline{C}_f, \tag{12}$$
$$\text{EDC}_9: \overline{C}_f < 5\overline{C}_r, \tag{13}$$

where $\overline{C}_f$ and $\overline{C}_r$ denote the mean of $C_f$ and $C_r$ over the simulation period.

– Yearly carbon pools growth rate is limited to 10 %.

$$\text{EDC}_{10-15}: \quad \overline{C}^n / \overline{C}^1 < 1 + \zeta (n-1)/10, \tag{14}$$

where for each pool $\overline{C}^i$ denotes the mean carbon pool size over year $i$ and the growth factor $\zeta$ is set to 1.

– Carbon pools are not expected to show rapid exponential decay; therefore, parameter sets are required to satisfy the condition that the half-life period of carbon pools is more than 3 years.

$$\text{EDC}_{16-21}: \quad \gamma < 3 \times 365/\log 2 \tag{15}$$

The trajectory of each carbon pool is approximated using an exponential decay curve $a + b \exp \gamma t$ where $a$, $b$ and $\gamma$ are the fitted exponential decay parameters and t the time variable, in days in this case.

– Carbon pools are expected to be within an order of magnitude of a steady-state attractor.

$$\text{EDC}_{22-29}: \quad C_0/10 < C^\infty < 10 C_0, \tag{16}$$

where for each of the carbon pools $C_s$, $C_l$, $C_w$ and $C_r$, $C_0$ denotes the initial state and $C^\infty$ denotes the steady-state attractor defined as

$$C_{\text{som}}^\infty = \frac{(f_{\text{wood}} + (f_{\text{fol}} + f_{\text{root}} + f_{\text{lab}})\, p_1)\,\overline{G}}{(p_1 + p_9)\, p_8 \exp \overline{T} p_{10}}, \tag{17}$$

$$C_{\text{lit}}^\infty = \frac{(f_{\text{fol}} + f_{\text{root}} + f_{\text{lab}})\,\overline{G}}{p_9 \exp \overline{T} p_{10}}, \tag{18}$$

$$C_{\text{wood}}^\infty = \frac{f_{\text{wood}}\overline{G}}{p_6}, \tag{19}$$

$$C_{\text{root}}^\infty = \frac{f_{\text{root}}\overline{G}}{p_7}, \tag{20}$$

where $\overline{G}$ denotes the mean gross primary production and $f_{\text{wood}}$, $f_{\text{som}}$ and $f_{\text{lit}}$ are given by

$$f_{\text{wood}} = 1 - f_{\text{auto}} - f_{\text{fol}} - f_{\text{lab}} - f_{\text{root}}, \tag{21}$$
$$f_{\text{som}} = f_{\text{wood}} + (f_{\text{root}} + f_{\text{lab}} + f_{\text{fol}})\, p_1/(p_1 + p_8), \tag{22}$$
$$f_{\text{lit}} = (f_{\text{root}} + f_{\text{lab}} + f_{\text{fol}}). \tag{23}$$

To the original EDCs, we found it useful to add the three following constraints:

$$\text{EDC}_{30}: \quad \text{LAI(summer)} < \alpha, \ \alpha > 0, \tag{24}$$
$$\text{EDC}_{31}: \quad \text{LAI(final day)} > 0, \tag{25}$$
$$\text{EDC}_{32,33}: \quad -\beta < E[\text{NEE}] < \beta, \ \beta > 0, \tag{26}$$

where $\alpha$ and $\beta$ are real constants that need to be adjusted, LAI(summer) denotes the modelled LAI during summer and LAI(final day) denotes the modelled LAI at the end of the model run. These new constraints guarantee that LAI and the mean NEE remain within realistic bounds.

Bloom and Williams (2015) demonstrated the efficiency of incorporating EDCs using a Monte Carlo method to improve parameter estimates and NEE predictions. We propose an approach to apply these extra constraints within a variational framework.
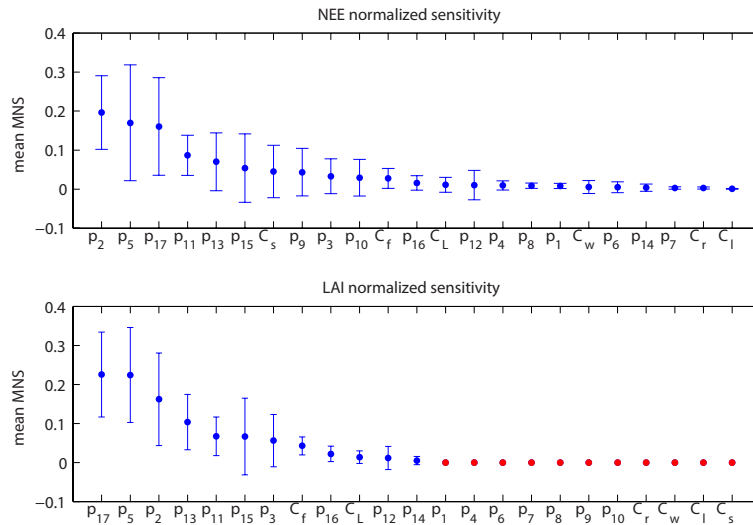
**Figure 2.** Mean normalised sensitivities (MNS): 100 parameter sets satisfying EDCs are sampled at the Morgan Monroe State Forest. Parameters are ranked in decreasing order according to their sensitivity, the blue dots represent the mean of the MNS (dimensionless quantity), the intervals represent $1\sigma$ error bars and the red dots correspond to null sensitivity.

## 2.3 Sensitivity analysis

Sensitivity analysis studies how the variations of the output $\boldsymbol{h}$ of a model can be attributed to variations of the input variables $x_i$. Such information is crucial for model design, inverse modelling and reduction of complex non-linear models. A global sensitivity analysis for DALEC was recently performed in Safta et al. (2015). Here we consider a local approach where first-order derivatives are used to build sensitivity indices that help us understand the influence of input variables on the output.

We denote $h_t$ as the function that maps $\boldsymbol{x} = \log(\boldsymbol{p}, \boldsymbol{C}^0)$ to the value of an output of the model (here LAI, NEE, GPP and RESP) at time $t$, and we denote the time series of the model output as $\boldsymbol{h} = (h_{t_1}, \ldots, h_{t_N})$. Following Zhu and Zhuang (2014), we consider the mean normalised sensitivity (MNS) defined as

$$s_i = E\left(|\frac{\partial \boldsymbol{h}}{\partial x_i}|\frac{\sigma_i}{\sigma_h}| / \sum_j |\frac{\partial \boldsymbol{h}}{\partial x_j}\frac{\sigma_j}{\sigma_h}|\right), \tag{27}$$

where $E(\cdot)$ denotes the average of the time series. The scalars $\sigma_i$ and $\sigma_h$ denote the parameter variance, set as 40 % of the parameter range, and the variance of the output respectively. The partial derivatives are computed using the adjoint derived using the method described in Giering and Kaminski (1998). The MNS $s_i$ is a dimensionless number that allows us to compare among parameters.

We consider the Morgan Monroe State Forest over a 12-year period. We sample 100 parameter sets satisfying the ecological constraints. For each parameter set, we compute the MNS for DALEC simulated mean fluxes LAI and NEE. In Fig. 2 parameters are ranked with respect to their mean

MNS. We see that for LAI only 12 out of the 23 variables are sensitive, namely $p_5$, $p_{17}$, $p_2$, $p_{13}$, $p_{11}$, $p_{15}$, $p_{16}$, $C_f$, $p_{12}$, $p_3$, $C_{\text{lab}}$ and $p_{14}$. Therefore, using LAI only in an inverse modelling experiment provides, at best, information about those twelve sensitive variables. For NEE we see that all variables are sensitive. Sensitivity analysis for GPP shows similar characteristics with LAI and so does RESP with NEE. For the four outputs under consideration (LAI, NEE, GPP and RESP) the most sensitive variables are the autotrophic respiration, $p_2$, the annual leaf loss fraction, $p_5$, the leaf mass per area, $p_{17}$, the fraction of GPP allocated to labile pool, $p_{13}$, the nitrogen use efficiency, $p_{11}$, and the leaf fall day $p_{15}$.

Here our focus is on the mean of the time series of DALEC fluxes (LAI, NEE) over a 12-year period. Finer analysis could be carried out by looking at seasonal aspects of the carbon cycle, identifying what variables are the most sensitive at certain times of the year, for example as studied in Safta et al. (2015).

## 3 Data assimilation

In this section we introduce concepts and methods that allow for close mathematical scrutiny of inverse problems and we present the variational method that we will apply in the following sections.

### 3.1 Ill-posed problem

A generic inverse problem consists of finding a $n$ dimensional state vector $\boldsymbol{x}$ such that

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{y}, \tag{28}$$

for a given $N$-dimensional observation vector $\boldsymbol{y}$, including random noise, and a given model $\boldsymbol{h}$. In the remainder of the paper the terms state vector, state variable, input variable and parameters will be used interchangeably to denote the vector $\boldsymbol{x}$ to be estimated using inverse methods and defined in the previous section as $\boldsymbol{x} = \log([\boldsymbol{p}, \boldsymbol{C}])^T$. The problem is well posed in the sense of Hadamard (1923) if the three following conditions hold: (1) there exists a solution, (2) the solution is unique and (3) the solution depends continuously on the input data. If at least one of these conditions is violated the problem is said to be ill-posed. The inverse problem (Eq. 28) is often ill-posed, and a regularisation method is required to replace the original problem with a well-posed problem. Solving Eq. (28) amounts to (1) constructing a solution $\boldsymbol{x}$, (2) assessing the validity of the solution and (3) characterising its uncertainty. Each inverse problem has its own features which need to be understood in order to characterise properly the solution and its uncertainty.

## 3.2 Bayesian inference: 4DVAR

Inverse problems are generally presented in a probabilistic framework where most methods can be expressed through a Bayesian formulation. The Bayesian approach provides a full characterisation of all possible solutions, their relative probabilities and uncertainties.

From Bayes' theorem, the probability density function (PDF) of the model state $\boldsymbol{x}$ given the set of observations $\boldsymbol{y}$, $p(\boldsymbol{x}|\boldsymbol{y})$, is given by

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}), \tag{29}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the PDF of the observations given $\boldsymbol{x}$ and $p(\boldsymbol{x})$ is the prior PDF of $\boldsymbol{x}$. A special case is given when $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x})$ are Gaussian PDF given by

$$p(\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \mathbf{B}^{-1}(\boldsymbol{x} - \boldsymbol{x}_0)\right], \tag{30}$$

and

$$p(\boldsymbol{y}|\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y})^T \mathbf{R}^{-1}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y})\right], \tag{31}$$

where $\mathbf{B}$ is the covariance matrix of the prior term $\boldsymbol{x}_0$, and $\mathbf{R}$ is the covariance matrix of the observation error. When the operator $\boldsymbol{h}$ is linear then the posterior PDF $p(\boldsymbol{x}|\boldsymbol{y})$ is Gaussian and thus fully characterised by its mean and covariance matrix. The mean is obtained by minimising the modulus of the log of the joint probability distribution, which is the cost function $J$ given by

$$J(\boldsymbol{x}) = J_0(\boldsymbol{x}) + J_y(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbf{B}}^2 + \frac{1}{2}\|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y}\|_{\mathbf{R}}^2. \tag{32}$$

Many methods can be considered to minimise this cost function. A Monte Carlo method is employed in Bloom and Williams (2015). Here we use a variational approach which

applies a gradient-based method where the gradient is given by

$$\nabla J = \mathbf{B}^{-1}(\boldsymbol{x} - \boldsymbol{x}_0) + \mathbf{H}^T \mathbf{R}^{-1}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{y}), \tag{33}$$

with $\mathbf{H}^T$ denoting the adjoint operator. The covariance matrix of the solution, $\mathbf{C}$, is given by the inverse of the Hessian of the cost function

$$\mathbf{C} = [\mathrm{Hess}(J)]^{-1} = \left[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}\right]^{-1}. \tag{34}$$

When the observation operator $\boldsymbol{h}$ is non-linear, the cost function $J$ can have multiple local minima and the posterior PDF may no longer be a Gaussian PDF. However, locally, the PDF $N(\tilde{\boldsymbol{x}}, \mathbf{C})$, where $\mathbf{C}$ is given by Eq. (34) evaluated at a minimum $\tilde{\boldsymbol{x}}$, provides a Gaussian approximation of the posterior PDF $p(\boldsymbol{x}|\boldsymbol{y})$.

The first term in the cost function (Eq. 32) is a regularisation term encoding the Gaussian prior $p(\boldsymbol{x})$. As we will show in the next sections the problem of assimilating Earth observations (LAI, GPP, NEE, RESP) into DALEC is a highly ill-posed problem and regularisation is required. The sensitivity analysis of Sect. 2.3 showed that LAI and GPP are not sensitive to all variables. Moreover, all observations streams show very low sensitivities to some variables. Therefore, as will be illustrated in Sect. 4.1, the solution (if any) is likely to be subject to large uncertainties. Apart from a couple of extensively studied sites, our prior knowledge about the variables is so far limited to their upper and lower bounds given in Table 1. As performed in Zhu and Zhuang (2014), it is a common practice to use this information to define a Gaussian prior $p(\boldsymbol{x}) \sim N(\boldsymbol{x}_0, \mathbf{B})$, where $\boldsymbol{x}_0$ is given by the centre of the variables ranges and $\mathbf{B}$ is the diagonal matrix whose diagonal elements are the squares of 40 % of the variables ranges. While using this kind of regularisation is necessary to ensure any solution at all when no better source of information is available, this introduces some biases in the solution. The EDCs introduced by Bloom and Williams (2015) provide new prior information about the variables. One of the purposes of this paper is to incorporate the EDCs as a regularisation term within 4DVAR. In the next section we propose a strategy to achieve this goal.

## 3.3 EDCs and 4DVAR

Incorporating the EDCs from an optimisation point of view can be easily performed by considering an inequality constraint optimisation problem where we aim at solving

$$\min_x J_y(\boldsymbol{x}) \text{ subject to } \boldsymbol{l} < \boldsymbol{x} < \boldsymbol{u} \text{ and } \boldsymbol{g}(\boldsymbol{x}) < \boldsymbol{0},$$

where $\boldsymbol{g}$ is the non-linear operator defining the EDCs described in Sect. 2.2, and $\boldsymbol{l}$ and $\boldsymbol{u}$ are the lower and upper bounds defined in Table 1. This approach provides an efficient, robust and quick strategy to find an acceptable solution; however, stability properties are not easily determined (see Roese-Koerner et al., 2012).

We are seeking a multivariate Gaussian distribution that would encode the EDCs. At a forest site, we start by sampling the parameter space to obtain an ensemble of 1000 parameter sets satisfying the EDCs; each parameter set $x$ is randomly created and required to satisfy $g(x) < 0$. We denote this ensemble by $\mathcal{X}_{EDCs}$. For most parameters, the sampling gives rise to undetermined PDFs which can certainly not be represented by Gaussian PDFs. However, upon inspecting the distribution $g(x)$, for all $x$ in $\mathcal{X}_{EDCs}$, we see that the distribution $\log(g(x))$ can be fairly accurately approximated by multivariate Gaussian PDFs $N(c, \Sigma)$, where $c$ denotes the mean of the distribution $\log(g(x))$ and $\Sigma$ denotes its covariance matrix. As an example, Fig. 3 shows the marginals $\log(g_4(x))$ and $\log(g_6(x))$, corresponding to the EDCs 4 and 6 respectively, together with a Gaussian fit.

Using Bayes' theorem we can then write

$$p(x|y, c) \propto p(y|x)p(c|x)p(x). \tag{35}$$

Finding a Gaussian approximation for $p(x|y, c)$ amounts then to minimising the cost function

$$J(x) = \frac{1}{2}\|h(x) - y\|_{\mathbf{R}}^2 + \frac{1}{2}\|\log(g(x)) - c\|_{\Sigma}^2$$
$$+ \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2. \tag{36}$$

The gradient of $J$ is given by

$$\nabla J(x) = \mathbf{H}^T \mathbf{R}^{-1}(h(x) - y) + \frac{1}{g(x)}\mathbf{G}^T \Sigma^{-1}(\log(g(x)) - c)$$
$$+ \mathbf{B}^{-1}(x - x_0),$$

and the Hessian of the cost function can be approximated by

$$\Theta = \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} - \frac{1}{(g(x))^2}\mathbf{G}\mathbf{G}^T \Sigma^{-1}(\log(g(x)) - c)$$
$$+ \frac{1}{(g(x))^2}\mathbf{G}^T \Sigma^{-1}\mathbf{G} + \mathbf{B}^{-1}, \tag{37}$$

evaluated at the minimiser $\widetilde{x}$. The operator $\mathbf{G}^T$ denotes the adjoint of the tangent linear model $\mathbf{G}$ whose key ingredient is given by the adjoint of DALECv2. The approximation of $p(x|y, c)$ is then given by the Gaussian distribution $N(\widetilde{x}, \Theta^{-1})$. In Sect. 5 we will perform experiments using real data to validate this approach.

# 4 Linear analysis

Considerable theoretical insights into the nature of the inverse problem, and the ill-posedness, can be obtained by studying a linearisation of the operator $h$. A first approximation to the inverse problem consists of finding a perturbation $z$ which best satisfies the linear equation
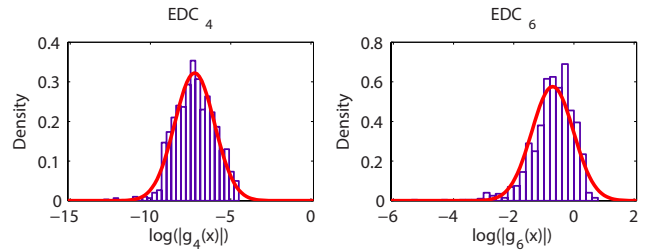
$$\mathbf{H}z = d, \tag{38}$$

**Figure 3.** Distribution and Gaussian fit for $EDC_4$ and $EDC_6$.

where $\mathbf{H}$ is the tangent linear operator for $h$ and $d$ is a perturbation of the observations. The linear operator $\mathbf{H}$ is commonly referred to as the *observability matrix* (see Johnson et al., 2005). The least squares formulation of this problem is to solve the optimisation problem

$$\min_z J(z) = \min_z \frac{1}{2}\|\mathbf{H}z - d\|^2. \tag{39}$$

The minimisation can be performed using an iterative method such as the conjugate gradient method, where the gradient is given by

$$\nabla J = \mathbf{H}^T(\mathbf{H}z - d). \tag{40}$$

The inverse Hessian of the cost function, $(\mathbf{H}^T \mathbf{H})^{-1}$, gives the covariance matrix of the least squares solution. In the next section we consider a direct solution method based on the singular value decomposition of the operator $\mathbf{H}$, which allows us to investigate the nature of the ill-posedness of the problem. We illustrate regularisation using a truncated singular value decomposition.

## 4.1 Singular value decomposition

We consider a singular value decomposition of $\mathbf{H}$ of the form

$$\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{41}$$

where $\mathbf{U}$ is a $N \times N$ unitary matrix, $\mathbf{V}$ is a $n \times n$ unitary matrix and $\mathbf{S}$ is the $N \times n$ diagonal matrix whose diagonal elements are the singular values $s_1 \geq \ldots \geq s_n \geq 0$. Using this decomposition, the solution $z_{LS}$ to Eq. (39) can be written as

$$z_{LS} = \mathbf{V}\mathbf{S}^\dagger \mathbf{U}^T y = \mathbf{H}^\dagger y. \tag{42}$$

The matrix $\mathbf{H}^\dagger = \mathbf{V}\mathbf{S}^\dagger \mathbf{U}^T$ is the *pseudo-inverse* of $\mathbf{H}$ where $\mathbf{S}^\dagger$ is the diagonal matrix obtained by transposing $\mathbf{S}$ and replacing the non-zero elements with their inverse $s_i^{-1}$. The covariance of the solution is given by

$$\mathrm{Cov}(z_{LS}) = \mathbf{H}^{\dagger^T}\mathbf{H}^\dagger. \tag{43}$$

Much can be learned about the stability of the solution (Eq. 42) by inspecting the singular values of $\mathbf{H}$. Assuming that $\mathbf{H}$ is full rank, it can be shown (see Golub and Van Loan,
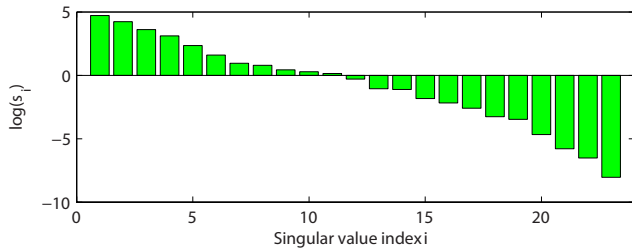
**Figure 4.** Singular values of the observability matrix for NEE (log scale).



**Figure 5.** Solution of the linearised inverse problem for NEE. The red points represent the observations, the red curve is the true trajectory, the green curve is the trajectory obtained using the unstable solution and the blue curve is obtained using the truncated singular value decomposition (TSVD) solution.

1996) that the relative error in the solution, defined as the left-hand side of the above inequality, is bounded by

$$\frac{\|z_{LS} - z_0\|}{\|z_0\|} \leq \kappa(\mathbf{H}) \frac{\|\boldsymbol{\epsilon}\|}{\|\boldsymbol{d}\|}, \tag{44}$$

where $\kappa(\mathbf{H})$ is the condition number of $\mathbf{H}$ defined as $\kappa(\mathbf{H}) = s_1/s_n$, $z_0$ denotes the truth (possibly unknown) and $\boldsymbol{\epsilon}$ represents observational noise. When the condition number is large the matrix is said to be ill-conditioned, the problem is ill-posed and the solution (Eq. 42) is unstable: small perturbations to the system can lead to very large perturbations in the solution.

### 4.2 Stability for NEE operator

As an example we consider the problem of assimilating NEE observations into DALECv2 to estimate model parameters and initial conditions at Morgan Monroe State Forest. We linearise Eq. (28) about a point $\boldsymbol{x}^*$ satisfying the EDCs, form the observability matrix $\mathbf{H}$ and compute its singular value decomposition. The singular values, shown in Fig. 4, reveal a condition number of the order of $10^5$.

For a signal-to-noise ratio, namely $\|\boldsymbol{\epsilon}\|/\|\boldsymbol{d}\|$, of magnitude 0.1, inequality (Eq. 44) gives an upper bound for the relative error in the solution of the order of $10^4$, which does not give much credit to the least squares solution. How sharp is this bound? Are we overestimating the error? To answer these questions we create a set of noisy observations with noise variance $\sigma = 0.1$ and we compute the solution (Eq. 42). The relative error for each component of the solution, $\eta_i$, and the variance $v_i$, are given in Table 2. Despite a relatively good match between the modelled NEE perturbations and the observations, as shown in Fig. 5, the results of Table 2 show very large relative errors and variances for most variables. Moreover, these results are in agreement with the results of REFLEX: parameters directly linked to foliage and GPP are better estimated than parameters related to allocation to and turnover of fine root/wood. The results of Table 2 reflect the sensitivity analysis shown in Fig. 2. The variables with respect to which NEE is the most (least) sensitive are the less (more) affected by the noise.

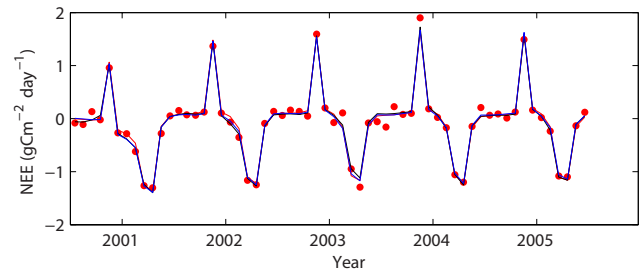To reduce the impact of observational noise on the solution, regularisation is required. The truncated singular value decomposition (TSVD) is a simple and popular method for regularisation. TSVD consists of truncating the pseudo-inverse in Eq. (42) in order to remove the smallest singular values, the most affected by the noise. The solution $z^{(k)}$ is then given by

$$z^{(k)} = \mathbf{V}_k \mathbf{S}_k^\dagger \mathbf{U}_k^T \boldsymbol{y} = \mathbf{H}_k^\dagger \boldsymbol{y}, \tag{45}$$

where $k$ is the truncation rank and where $\mathbf{S}_k$, $\mathbf{U}_k$ and $\mathbf{V}_k$ are the rectangular matrices formed by the first $k$ columns of $\mathbf{S}$, $\mathbf{U}$ and $\mathbf{V}$. The covariance of the solution is given by

$$\text{Cov}\left(z^{(k)}\right) = \mathbf{V}_k \left(\mathbf{S}_k^\dagger\right)^{-2} \mathbf{V}_k^T. \tag{46}$$

The truncation rank $k$ can be chosen using the L-curve method. The L curve is a log–log plot of the norm of the solution $\|z^{(k)}\|$ against the norm of the residual $\|\mathbf{H}z^{(k)} - \boldsymbol{d}\|$ parametrised by the regularisation parameter $k$. The optimal parameter corresponds to the point of maximum curvature of the L curve. Further details on the L-curve method can be found in Hansen and O'Leary (1993).

In our example with NEE we use Hansen's regularisation tools (see Hansen, 2007) to perform the TSVD method. The truncation rank obtained using the L-curve method is $k = 7$. The last three columns of Table 2, presenting the TSVD solution, the relative error of each components and the variances, can be compared with the unstable solution results. Whereas the relative errors in the unstable solution range from $5.3 \times 10^{-2}$ to $3.8 \times 10^4$, the relative errors in the regularised solution range from $5.3 \times 10^{-2}$ to 5.1. We see that TSVD has the effect of keeping small the variables that cannot be estimated correctly. As previously stated the results of the regularisation can be related to the sensitivity analysis depicted in Fig. 2: TSVD prevents the variables with respect to which NEE is the least sensitive from growing unbounded.

In the next section we consider the concept of a resolution matrix, which allows for finer analysis of the solution of the linear problem.

**Table 2.** Results of the linear inverse problem showing (1) the solution components for the least squares solution $z_{LS}$ together with their relative error $\eta_i$ (dimensionless quantity) and variance $v_i$ and (2) the solution components for the TSVD solution $z^{(k)}$ together with their relative error $\eta_i^{(k)}$ and variance $v_i^{(k)}$.

|            | $x^*$    | $z^*$   | $z_{LS}$  | $\eta_i$  | $v_i$          | $z^{(k)}$ | $\eta_i^{(k)}$ | $v_i^{(k)}$ |
|------------|----------|---------|-----------|-----------|----------------|-----------|----------------|-------------|
| $p_1$      | −6.984   | −0.070  | 3.715     | 54.190    | 19 182.5715    | 0.004     | 1.052          | 0.0001      |
| $p_2$      | −1.114   | −0.011  | 0.342     | 31.715    | 23.2871        | 0.003     | 1.242          | 0.0005      |
| $p_3$      | −3.480   | −0.035  | −2.690    | 76.285    | 7384.9940      | 0.001     | 1.022          | 0.0000      |
| $p_4$      | −2.745   | −0.027  | 3.389     | 124.470   | 82 380.0339    | −0.000    | 1.000          | 0.0000      |
| $p_5$      | 0.086    | 0.001   | 0.048     | 54.082    | 0.6575         | −0.004    | 5.139          | 0.0009      |
| $p_6$      | −8.776   | −0.088  | 67.445    | 769.477   | 1 581 516.7404 | 0.000     | 1.001          | 0.0000      |
| $p_7$      | −5.265   | −0.053  | −0.999    | 17.970    | 57.1478        | −0.005    | 0.900          | 0.0001      |
| $p_8$      | −6.640   | −0.066  | −0.344    | 4.176     | 7981.3944      | −0.016    | 0.757          | 0.0013      |
| $p_9$      | −10.292  | −0.103  | 133.504   | 1298.187  | 494 620.7529   | −0.006    | 0.946          | 0.0002      |
| $p_{10}$   | −3.035   | −0.030  | −0.003    | 0.889     | 2.6980         | −0.011    | 0.632          | 0.0008      |
| $p_{11}$   | 3.539    | 0.035   | 1.075     | 29.370    | 79.5237        | −0.003    | 1.083          | 0.0003      |
| $p_{12}$   | 4.736    | 0.047   | 0.045     | 0.053     | 0.0256         | 0.044     | 0.080          | 0.0003      |
| $p_{13}$   | −0.772   | −0.008  | 1.042     | 135.879   | 8676.9499      | −0.028    | 2.616          | 0.0033      |
| $p_{14}$   | 3.261    | 0.033   | 0.072     | 1.196     | 4.3971         | 0.004     | 0.866          | 0.0001      |
| $p_{15}$   | 5.533    | 0.055   | 0.084     | 0.515     | 0.0548         | 0.058     | 0.053          | 0.0003      |
| $p_{16}$   | 4.082    | 0.041   | 0.092     | 1.265     | 1.2521         | 0.004     | 0.904          | 0.0001      |
| $p_{17}$   | 5.178    | 0.052   | 1.631     | 30.497    | 8160.2559      | 0.000     | 0.997          | 0.0005      |
| $C_{lab}$  | 6.237    | 0.062   | 1.002     | 15.073    | 8237.5716      | 0.019     | 0.697          | 0.0020      |
| $C_f$      | 4.073    | 0.041   | 1.348     | 32.090    | 8070.9246      | 0.005     | 0.888          | 0.0001      |
| $C_r$      | 6.858    | 0.069   | 1.788     | 25.067    | 8300.6094      | −0.012    | 1.170          | 0.0008      |
| $C_w$      | 8.341    | 0.083   | −318.175  | 3815.484  | 7 436 253.3370 | 0.000     | 0.999          | 0.0000      |
| $C_l$      | 5.961    | 0.060   | 0.568     | 8.532     | 8550.3479      | −0.006    | 1.097          | 0.0002      |
| $C_s$      | 8.956    | 0.090   | −134.334  | 1500.869  | 483 025.4281   | −0.006    | 1.064          | 0.0002      |

## 4.3 Resolution matrix

As suggested by Eqs. (42) and (45), finding a solution $z$ amounts to constructing a *generalised inverse* $\mathbf{H}^g$ such that formally

$$z = \mathbf{H}^g d. \qquad (47)$$

The generalised inverse is the operator representing any method, direct or iterative, used to solve the linear inverse problem, with or without any kind of regularisation. In the previous section we considered two examples of generalised inverse, the pseudo-inverse and the truncated inverse obtained using TSVD. The generalised inverse can be used to define operators which directly address the conditions for well-posedness for the linearised problem. Assuming a true state $z^*$ exists, possibly unknown, using Eqs. (38) and (47) we can then define an operator $\mathbf{N}$ called the *model resolution matrix* which relates the solution $z$ to the true state

$$z = \mathbf{H}^g \mathbf{H} z^* = \mathbf{N} z^*. \qquad (48)$$

This matrix provides a practical tool to analyse the resolution power of an inverse method, that is, its ability to retrieve the true state, with or without using any regularisation method: the closer $\mathbf{N}$ is to the identity, the better the resolution. Moreover, the trace of the matrix defines a natural notion of *information content* (IC). Similarly a *data resolution matrix* can

be defined to study how well data can be reconstructed and its diagonal elements naturally define a notion of *data importance*. For the two examples of generalised inverse presented in the previous section we obtain the following resolution matrices:

$$\mathbf{N} = \mathbf{H}^\dagger \mathbf{H}, \qquad (49)$$

for the pseudo-inverse and

$$\mathbf{N} = \mathbf{V}_k \mathbf{V}_k^T, \qquad (50)$$

for the truncated pseudo-inverse. In the first case the IC equals the number of non-zero singular values, in the second case the IC equals the truncation rank $k$. An in-depth theoretical and practical analysis of these concepts and those introduced in the remainder of this section can be found in Menke (1984).

While the model resolution matrix allows us to see how a solution strategy maps the true state variables to the solution of the inverse problem, and to see how well and how independently the state variables can be recovered, one also needs to assess the uncertainty of the solution. This can be studied using the so-called *unit covariance matrix*, $\mathbf{C}$, defined using the generalised inverse as
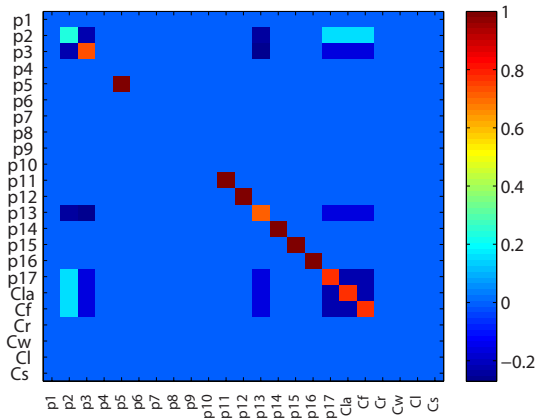
$$\mathbf{C} = \mathbf{H}^{g^T} \mathbf{H}^g. \qquad (51)$$

**Figure 6.** Model resolution matrix for the LAI operator.



**Figure 7.** Diagonal elements (log scale) of the unit covariance matrix for the LAI operator: using the pseudo-inverse shown in green and TSVD shown in yellow.

By characterising the degree of error amplification that occurs in the mapping from the true state to the solution of the inverse problem, the unit covariance matrix is a crucial object for studying the stability of the solution with respect to observational noise. The unit covariance matrix defined by Eq. (51) agrees with the covariance matrices given in the previous section by Eq. (43) for the pseudo-inverse, and by Eq. (46) when TSVD is applied.

## 4.4 Resolution for LAI operator

We now study the model resolution matrix for the LAI observation operator at Morgan Monroe State Forest. In the first instance we will demonstrate the resolution power of the LAI signal without regularisation using the pseudo-inverse as generalised inverse first, and then apply TSVD to show how using the truncated pseudo-inverse affects resolution. In a second case we will study the added value of the EDCs in terms of resolution.

As previously, we linearise Eq. (28) about the point $x^*$ given in Table 2. The trace of the resolution matrix obtained using the pseudo-inverse as generalised inverse is 10, and this means that 10 independent variables can be estimated using LAI. These independent variables are not the variables in which the system is expressed, but a linear transformation can be found to express the system in terms of the independent variables. Figure 6 shows the model resolution matrix for LAI. As shown in Sect. 2.3 with the sensitivity analysis, 11 out of the 23 variables are not sensitive to LAI, and this can be seen in the resolution matrix by the diagonal terms which are zero, represented in blue. In contrast the diagonal elements corresponding to sensitive variables have positive values, represented by colours ranging from light blue to red. Figure 6 also shows that whereas $p_5$, $p_{11}$, $p_{12}$, $p_{14}$, $p_{15}$ and $p_{16}$ are perfectly resolved (the corresponding elements are coloured brown or dark red), there exist linear combinations between the remaining sensitive variables, which explains
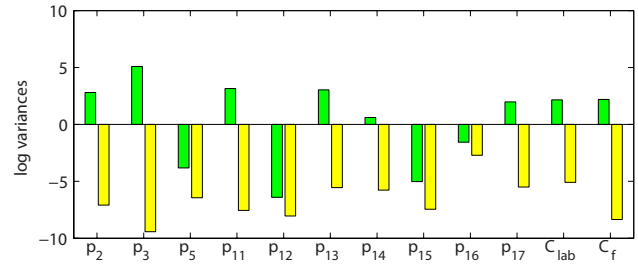
why only 10 independent variables can be estimated from the 12 sensitive variables.

For the study of the unit covariance matrix we restrict ourselves to the sensitive variables. This amounts to removing the columns corresponding to the non-sensitive variables, containing only null elements, from the observability matrix. The dependency of the solution on observational noise can be studied by looking at Fig. 7, where the diagonal elements of the unit covariance matrix, corresponding to the variance of each element of the solution obtained using the pseudo-inverse, are represented in log scale. Except for $p_5$, $p_{12}$, $p_{15}$ and $p_{16}$, all variances are shown to be large.

As previously, we illustrate a simple regularisation strategy, TSVD, and show its effects on both resolution and stability. Figure 8 shows the resolution matrix for LAI with optimal truncation rank $k = 6$. The IC decreases to 6. We see that whereas $p_5$, $p_{12}$, $p_{15}$ and $p_{16}$ remain almost perfectly resolved, $p_{13}$, $p_{17}$ and $C_{lab}$ are only partially resolved and the remaining variables are not resolved properly. Figure 7 shows the corresponding diagonal elements of the unit covariance matrix, from which we see that the variances have been drastically reduced. This example shows how regularisation ensures stability at the price of losing resolution.

We now consider the effect of incorporating the static EDCs into the variational framework in terms of resolution. The static EDCs are given by the first seven EDCs, the linear problem is then given by

$$\begin{bmatrix} \mathbf{H} \\ \widetilde{\mathbf{G}} \end{bmatrix} z = \begin{pmatrix} d \\ f \end{pmatrix},$$ (52)

where

$$\widetilde{\mathbf{G}} = g(x^*)^{-1} \Sigma^{-1/2} \mathbf{G},$$ (53)

with $\Sigma^{-1/2}$ the inverse of the symmetric square root of the covariance matrix $\Sigma$, defined in Sect. 3.3, restricted to the first seven components. The static EDCs depend only on 13 out of the 23 variables, namely $p_1$ to $p_{10}$, $p_{12}$, $p_{13}$ and $p_{15}$. This can be seen on the matrix $\mathbf{G}$ where the columns corresponding to the remaining variables are null. Together with LAI observations, whose sensitive variables are represented in Fig. 2, we therefore have 19 sensitive variables. The
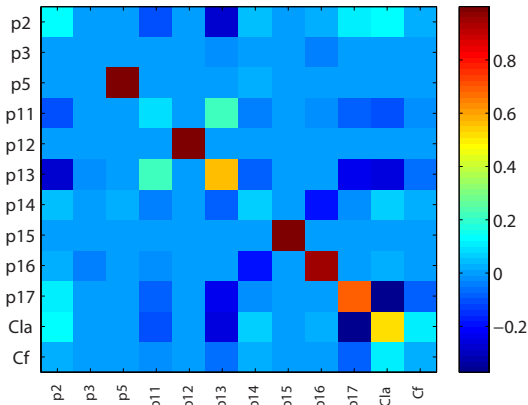
**Figure 8.** Model resolution matrix for the LAI operator using TSVD with truncation rank $k = 5$.
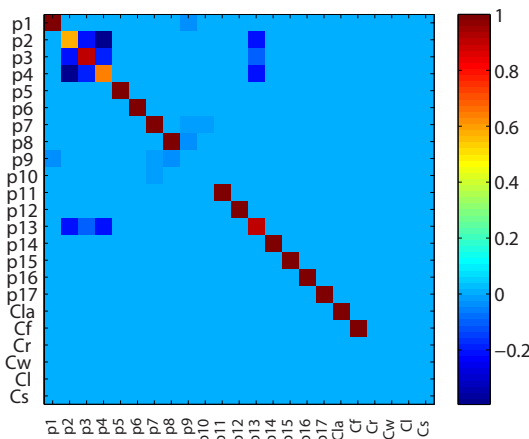


**Figure 9.** Model resolution matrix for LAI and static EDCs as defined by Eq. (52).

model resolution matrix corresponding to the operator on the left-hand side of Eq. (52), obtained using the pseudo-inverse, is depicted in Fig. 9. The trace of the model resolution matrix gives an IC of 16, 13 variables are perfectly resolved and 4 variables show linear dependencies ($p_2$, $p_3$, $p_4$ and $p_{13}$). However, although $p_9$ and $p_{10}$ are sensitive variables, they do not appear to be resolved at all: inspecting the linear operator **G** shows that the non-zero components corresponding to $p_9$ and $p_{10}$ are several order of magnitude smaller than the other components.

This example shows clearly the benefit of introducing the static EDCs to help estimate poorly constrained or otherwise undetermined components.

## 5 Experiments at AmeriFlux sites

We now consider a real experiment at the Morgan Monroe State Forest. At this AmeriFlux site, 12 years of MODIS LAI monthly mean observations from 2001 to 2013, NEE, GPP

**Table 3.** Experiment set up summary: in Exp. 1 we use LAI and bounds constraints (BDS), in Exp. 2 we use LAI, NEE and BDS and so on.

|        | LAI | NEE | GPP | RESP | BDS | EDCs |
|--------|-----|-----|-----|------|-----|------|
| Exp. 1 | x   |     |     |      | x   |      |
| Exp. 2 | x   | x   |     |      | x   |      |
| Exp. 3 | x   | x   | x   | x    | x   |      |
| Exp. 4 | x   |     |     |      | x   | x    |
| Exp. 5 | x   | x   |     |      | x   | x    |
| Exp. 6 | x   | x   | x   | x    | x   | x    |

and thus RESP observations from 2001 to 2005 are available. Our goal is to study two different aspects. The first one is the impact of using multiple data streams: how does it affect uncertainty of the predicted fluxes and how well do we predict non-observed fluxes? The second one is to use the static EDCs and to assess their utility in constraining poorly sensitive variables.

When including all terms the cost function, $J_{\text{TOT}}$, becomes

$$
\begin{aligned}
J_{\text{TOT}}(\boldsymbol{x}) = {} & \frac{\lambda_{\text{L}}}{2} \|\boldsymbol{h}_{\text{L}}(\boldsymbol{x}) - \boldsymbol{y}_{\text{L}}\|^2 + \frac{\lambda_{\text{N}}}{2} \|\boldsymbol{h}_{\text{N}}(\boldsymbol{x}) - \boldsymbol{y}_{\text{N}}\|^2 \\
& + \frac{\lambda_{\text{G}}}{2} \|\boldsymbol{h}_{\text{G}}(\boldsymbol{x}) - \boldsymbol{y}_{\text{G}}\|^2 + \frac{\lambda_{\text{R}}}{2} \|\boldsymbol{h}_{\text{R}}(\boldsymbol{x}) - \boldsymbol{y}_{\text{R}}\|^2 \\
& + \frac{\lambda_c}{2} \|\log(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{c}\|_{\Sigma}^2 + \frac{\lambda_0}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\text{B}}^2 \\
= {} & J_{\text{L}} + J_{\text{N}} + J_{\text{G}} + J_{\text{R}} + J_c + J_0,
\end{aligned}
$$

where subscripts L, N, G and R stand for LAI, NEE, GPP and RESP respectively. The vectors $\boldsymbol{y}_{\text{L}}$, $\boldsymbol{y}_{\text{N}}$, $\boldsymbol{y}_{\text{G}}$ and $\boldsymbol{y}_{\text{R}}$ represent the observation vectors for LAI, NEE, GPP and RESP respectively. The scalars $\lambda_{\text{L}}$, $\lambda_{\text{N}}$, $\lambda_{\text{G}}$ and $\lambda_{\text{R}}$ take the value 0 or 1 depending on whether or not the corresponding data stream is included in the experiment. The scalar $\lambda_c$ takes the value 0 or 1 depending on whether we include the EDCs and $\lambda_0$ takes the value 1.

We perform six experiments summarised in Table 3. In experiment (Exp.) 1, we use only LAI observations and bounds constraints so that in the cost function $J_{\text{TOT}}$ we set $\lambda_{\text{L}} = 1$ and $\lambda_0 = 1$, and the other $\lambda$s are set to zero. For Exp. 2, we use LAI and NEE observations, that is, we set $\lambda_{\text{L}} = 1$, $\lambda_{\text{N}} = 1$ and $\lambda_0 = 1$; the other $\lambda$s are set to zero. We proceed similarly for the remaining experiments. Here we assimilate all data streams simultaneously; it is not our intention to question what method best accommodates multiple data streams. MacBean et al. (2016) addresses this question using a simple C cycle model. Moreover, we choose to assume the same statistical error for all data streams and set their error covariance matrix equal to the identity. To avoid being trapped at meaningless local minima, the experiments are performed multiple times using different initialisation parameter sets and results for the best candidate only are reported.

The results of the experiments are presented in Table 4, where each element of $J_{\text{TOT}}$ is given for all experiments, and
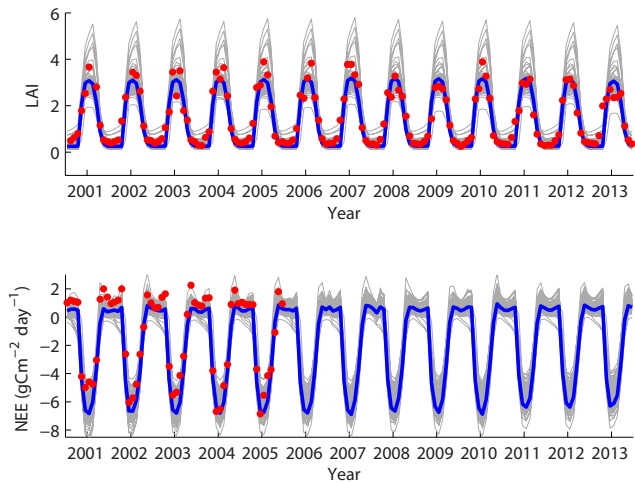
**Figure 10.** DALECv2 monthly estimates for LAI and NEE at Morgan Monroe State Forest. The red dots are the observations, the blue trajectories are obtained using the 4DVAR analysis, and the grey trajectories are ensemble runs obtained from a 95 % confidence sample of the posterior PDF.
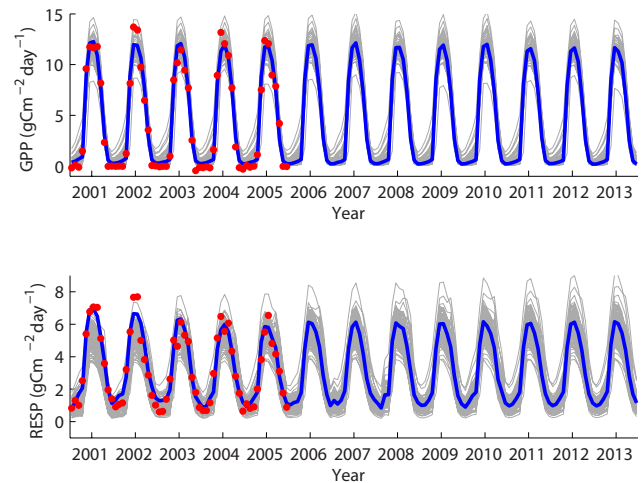


**Figure 11.** DALECv2 monthly estimates for GPP and RESP at Morgan Monroe State Forest. The red dots are the observations, the blue trajectories are obtained using the 4DVAR analysis, the grey trajectories are ensemble runs obtained from a 95 % confidence sample of the posterior PDF.

in Table 5, where the solution components and their variance are presented for all experiments. Results of Table 4 show that $J_L$ is the smallest in Exp. 1 when LAI only is used. In Exp. 2, when adding NEE we see that $J_N$ decreases from 109.012 in Exp. 1 to 15.263, and $J_G$ slightly decreases as compared to Exp. 1, but $J_R$ increases instead. In Exp. 3 we see that all costs drastically decrease compared to their initial values. Going from Exp. 1 to Exp. 3, $J_0$ slightly increases; adding more data streams constrains more parameters, and the parameters shift from their prior value which may cause $J_0$ to increase. Similar observations can be made for Exp. 4 to Exp. 6; moreover, we see that including the EDCs only slightly affects the costs. A reason for this might be that EDCs help constrain the less sensitive parameters for which the costs are less sensitive, as suggested by the sensitivity analysis depicted in Fig. 2. To see the effect of the EDCs we need to look at Table 5, which details the solution components together with their relative variance defined by the ratio of the variance by the parameter range. In Exp. 1 we see that the variables with the smallest relative variance are the most sensitive parameters as illustrated in Fig. 2: $p_2$, $p_5$, $p_{10}$, $p_{12}$, $p_{14}$, $p_{15}$, $p_{16}$ and $p_{17}$. We recall that the sensitivity analysis of Sect. 2.3 was performed by averaging sensitivities for an ensemble of initial parameter sets; therefore, the ranking shown in Fig. 2 may not be reflected in the relative variances. As we include NEE in Exp. 2 we see that most relative variances decrease, especially for $p_8$, $p_9$, $p_{10}$, $p_{13}$, $C_{lab}$, $C_f$ and $C_l$. The only variable whose relative variance increases is $p_{14}$, but as shown in Fig. 2, $p_{14}$ has very low sensitivity. In Exp. 3 most relative variances decrease. The values are still large though for $p_1$, $p_3$, $p_4$, $p_6$, $p_9$, $C_r$ and $C_l$. Again, similar features can be observed for Exp. 4 to Exp. 6,

but a clear improvement can be seen for most variables except for $C_r$ which is not constrained by the first seven EDCs. Finally, the last column of Table 4 shows the computation time for each experiment. As expected we see that the more observation streams we consider, the longer the experiment takes to run, and incorporating the EDCs increases computation time. However, we stress that these figures are several orders of magnitude less than the time required to perform the same experiments using the current gold standard MCMC approach used in Bloom and Williams (2015).

Figures 10 and 11 show the predicted fluxes for LAI, NEE, GPP and RESP for the result of Exp. 6. We can see good agreement between modelled fluxes and observations. The uncertainty of the predicted fluxes is evaluated by modelling an ensemble of trajectories from a 95 % ellipsoid of the posterior truncated Gaussian distribution. These trajectories are represented as grey curves in Figs. 10 and 11. Figure 12 shows the posterior parameter distribution marginals for $p_1$, $p_7$, $p_{16}$ and $C_f$ for Exp. 6, illustrating the four different cases where most of the marginal is contained in the parameter range for $p_{16}$; the marginal is truncated on the left or the right for $p_7$ and $C_f$ and truncated on both sides for $p_1$.

## 6 DALEC-SP

In the previous section we used EDCs within 4DVAR and showed their advantage in reducing drastically the uncertainty of otherwise undetermined variables. However, we only included the static EDCs which do not require a model run. As including more EDCs often leads to convergence issues, the solution and its uncertainty become subject to caution.

**Table 4.** Costs for the results of the inverse modelling experiments. The last column reports the computation time in seconds for the experiment.

|          | $J_L$    | $J_N$    | $J_G$     | $J_R$    | $J_0$  | $J_c$  | Time (s) |
|----------|----------|----------|-----------|----------|--------|--------|----------|
| $x_{init}$ | 179.525 | 353.229  | 1265.556  | 419.696  | 0.003  | 7.157  | 0.000    |
| Exp. 1   | 14.083   | 109.012  | 153.475   | 45.415   | 0.017  | 2.498  | 2.722    |
| Exp. 2   | 19.188   | 15.263   | 145.349   | 131.963  | 0.018  | 3.704  | 7.541    |
| Exp. 3   | 25.089   | 16.737   | 36.155    | 17.842   | 0.020  | 4.643  | 5.886    |
| Exp. 4   | 14.083   | 107.420  | 152.908   | 45.480   | 0.016  | 2.498  | 5.012    |
| Exp. 5   | 19.193   | 15.262   | 145.254   | 131.878  | 0.018  | 3.701  | 9.045    |
| Exp. 6   | 25.059   | 16.699   | 36.143    | 17.826   | 0.019  | 4.642  | 8.215    |
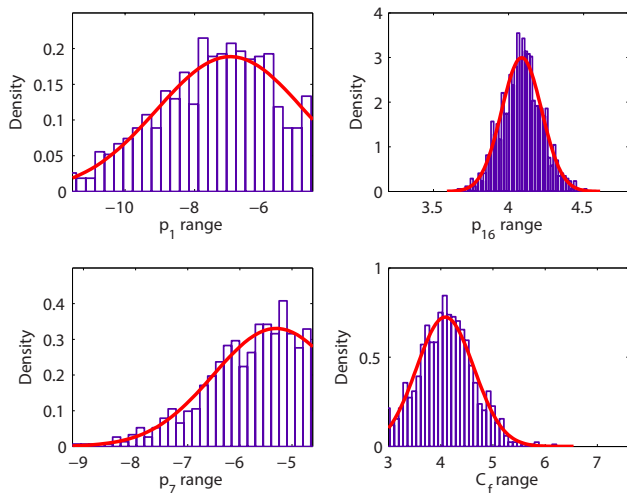


**Figure 12.** Posterior parameter distributions for parameters $p_1$, $p_7$, $p_{16}$ and $C_f$ for Exp. 6. For each plot the limits of the abscissa correspond to the parameter range. The red curve is the Gaussian posterior distribution and the blue bars represent the sample used to produce the grey trajectories in Figs. 10 and 11.

As shown in Chuter et al. (2015) for the previous DALEC evergreen and deciduous models, the evolution of the carbon pools for DALECv2 show a tipping point which depends on the parameters $p_1$ to $p_{17}$. Given a set of parameters, **p**, the fast carbon pools $C_{lab}$, $C_f$, $C_r$ and $C_l$ grow or decay rapidly to an equilibrium state. This equilibrium is either zero and the forest dies out or a pseudo-periodical seasonal cycle as shown in Fig. 13 for $C_f$. Moreover, there exists a limit value below which any initial condition leads to the zero equilibrium and above which the equilibrium is a strictly positive pseudo-periodical seasonal cycle.

Here we consider ecosystems with no recent major disturbance, where the fast carbon pools are expected to be close to their pseudo-periodical cycle. To model these ecosystems, one can either restrict the parameter space by using the dynamic EDCs, or we can introduce a spin-up period during which the carbon pools reach their attractor. Given parameters $p_1$ to $p_{17}$ and initial values for $C_w$ and $C_s$ a first run of DALECv2 is performed to obtain a state which is closer

to a pseudo-periodical cycle for the fast carbon pools. The steady-state trajectories are then used to initialise the fast carbon pools. For this DALECv2 "spin-up" model, DALEC-SP, the state variable is therefore formed of the 17 parameters, $p_1, \ldots, p_{17}$, and the initial conditions for $C_w$ and $C_s$.

DALEC-SP offers several advantages: some of the EDCs such as those controlling the growth and the half-life period of carbon pools are almost automatically satisfied. This greatly reduces the time required to generate the PDF $p(c|x)$. Moreover, as the sensitivity analysis and the resolution matrices showed, the fast carbon pools are variables that are not highly sensitive to the signals that we observe, and therefore reducing the number of variables by removing the fast carbon pools is likely to improve the overall conditioning of the inverse problem. Reproducing experiments 1 to 6 using DALEC-SP shows similar results to those with DALECv2.

## 7 Discussion

To our knowledge, this paper presents the first application of variational methods for an inverse modelling experiment using DALEC. Over the last 15 years many studies have validated the use of DALEC together with various types of data streams to infer ecological parameters at the site level. However, first ensemble Kalman filter and then Monte Carlo methods were privileged. At the same time 4DVAR has been successfully used at the global scale to constrain ecosystem parameters in carbon cycle data assimilation system (CC-DAS). In Rayner et al. (2005), the Biosphere Energy Transfer Hydrology model (BETHY) is coupled with the transport model TM2, and satellite observations of photosynthetically active radiation and atmospheric $CO_2$ concentration observations are used to optimise model parameters. In this context Kemp et al. (2014) investigated how to constrain the 4DVAR problem in CCDAS through a number of different methods: using constrained optimisation, adding a penalty term and applying parameter transformations. They concluded that using parameter transformations give the best results. In our study the three methods were investigated: Gaussian anamorphosis where priors based on the distribution of parameters satisfying the EDCs were considered, constrained optimisa-

**Table 5.** Results of the inverse modelling experiments. The solution components together with their relative variance, in brackets, are given for each experiment. The vector $x_{init}$ is the randomly chosen parameter set satisfying the EDCs that initialises the minimisation routine.

| | $x_{init}$ | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 5 | Exp. 6 |
|---|---|---|---|---|---|---|---|
| $p_1$ | −5.172 | −8.059 (1.727) | −8.248 (1.471) | −8.282 (1.021) | −5.954 (0.112) | −5.901 (0.075) | −6.499 (0.057) |
| $p_2$ | −0.947 | −0.885 (0.207) | −1.106 (0.120) | −1.085 (0.030) | −0.848 (0.171) | −0.982 (0.138) | −0.984 (0.020) |
| $p_3$ | −4.318 | −2.673 (0.955) | −2.944 (0.849) | −3.376 (0.894) | −3.073 (0.954) | −4.603 (0.973) | −3.510 (0.895) |
| $p_4$ | −1.493 | −2.649 (0.978) | −2.813 (0.961) | −2.692 (0.936) | −1.589 (0.155) | −1.386 (0.091) | −1.476 (0.095) |
| $p_5$ | 1.123 | 0.117 (0.003) | 0.153 (0.002) | 0.085 (0.000) | 0.135 (0.002) | 0.010 (0.000) | 0.090 (0.000) |
| $p_6$ | −7.959 | −8.752 (0.922) | −8.870 (0.911) | −8.707 (0.919) | −6.910 (0.144) | −7.933 (0.886) | −8.330 (0.883) |
| $p_7$ | −7.432 | −6.908 (1.151) | −6.373 (0.941) | −5.064 (0.224) | −7.336 (0.225) | −7.015 (0.207) | −6.241 (0.320) |
| $p_8$ | −5.281 | −6.908 (1.151) | −6.906 (0.316) | −6.522 (0.078) | −5.768 (0.107) | −4.981 (0.049) | −6.144 (0.034) |
| $p_9$ | −16.012 | −11.513 (2.303) | −10.075 (0.995) | −11.411 (1.514) | −10.563 (1.141) | −15.973 (2.298) | −11.043 (0.848) |
| $p_{10}$ | −3.041 | −3.272 (0.373) | −3.296 (0.085) | −3.036 (0.055) | −3.255 (0.371) | −2.689 (0.077) | −3.037 (0.051) |
| $p_{11}$ | 2.792 | 3.829 (0.540) | 4.026 (0.163) | 3.542 (0.003) | 3.958 (0.458) | 3.573 (0.120) | 3.548 (0.003) |
| $p_{12}$ | 3.549 | 4.626 (0.002) | 4.739 (0.000) | 4.735 (0.000) | 4.625 (0.002) | 4.663 (0.001) | 4.736 (0.000) |
| $p_{13}$ | −1.768 | −0.693 (0.130) | −0.996 (0.067) | −0.795 (0.046) | −0.930 (0.077) | −0.761 (0.026) | −0.813 (0.033) |
| $p_{14}$ | 3.343 | 4.013 (0.034) | 3.291 (0.123) | 3.292 (0.052) | 3.968 (0.030) | 3.762 (0.009) | 3.248 (0.077) |
| $p_{15}$ | 5.656 | 5.512 (0.001) | 5.528 (0.001) | 5.531 (0.000) | 5.518 (0.001) | 5.626 (0.000) | 5.533 (0.000) |
| $p_{16}$ | 4.529 | 4.115 (0.068) | 3.993 (0.025) | 4.095 (0.011) | 4.050 (0.063) | 4.463 (0.009) | 4.100 (0.010) |
| $p_{17}$ | 5.351 | 5.289 (0.213) | 5.278 (0.198) | 5.138 (0.165) | 5.129 (0.180) | 5.051 (0.104) | 5.082 (0.106) |
| $C_{lab}$ | 3.979 | 5.950 (0.115) | 6.031 (0.059) | 6.187 (0.040) | 5.792 (0.103) | 5.026 (0.143) | 6.106 (0.027) |
| $C_f$ | 5.389 | 4.677 (0.282) | 4.868 (0.068) | 4.038 (0.066) | 4.542 (0.263) | 5.806 (0.062) | 4.152 (0.043) |
| $C_w$ | 7.045 | 5.298 (1.151) | 5.829 (0.900) | 6.520 (0.096) | 5.329 (1.149) | 7.165 (0.265) | 7.093 (0.232) |
| $C_r$ | 9.753 | 8.406 (1.554) | 8.188 (1.533) | 8.318 (1.544) | 8.453 (1.541) | 9.612 (1.553) | 8.114 (1.531) |
| $C_l$ | 3.992 | 5.298 (1.151) | 7.307 (0.300) | 6.226 (0.161) | 5.354 (1.141) | 4.534 (0.438) | 6.015 (0.089) |
| $C_s$ | 9.721 | 8.406 (1.900) | 9.546 (1.188) | 8.603 (1.633) | 8.889 (1.528) | 9.615 (1.899) | 8.448 (1.559) |
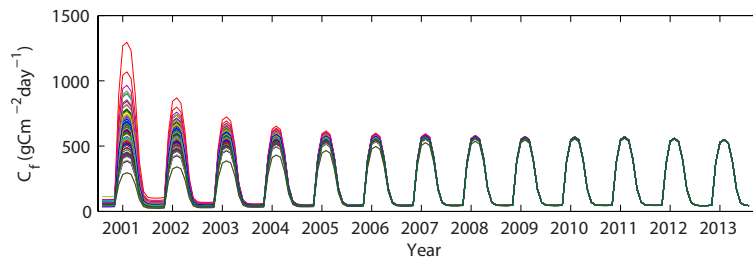


**Figure 13.** Pseudo-periodical seasonal cycle for DALECv2. Using a given set of parameters and initial values for $C_w$ and $C_s$, 100 DALECv2 runs are performed using random initial values for $C_{lab}$, $C_f$, $C_r$, $C_l$. The plot shows the 100 trajectories for $C_f$.

tion as stated in Sect. 3.3 and adding a penalty term to account for the EDCs. The latter solution which is the main interest of this publication was found to be the most successful in our case.

The complexity of global-scale experiments still limit the application of fully non-linear methods such as MCMC. In Ziehn et al. (2012) a comparison between the MCMC Metropolis–Hastings approach and 4DVAR for the BETHY-TM2 CCDAS framework is performed. This study reports a computation time of less than 1 h for the variational method and about 8 months for the overall MCMC computation. For our setting, DALECv2 site-based experiment, the complexity is relatively small and a MCMC approach is affordable. Used in Bloom and Williams (2015), the MCMC approach for DALEC is studied in detail in Safta et al. (2015), and the resulting parameter distributions suggest that 4DVAR and the

inherent Gaussian approximation provide a reasonable posterior distribution.

As with most variational methods, the analysis and application presented in this paper rely heavily on the possibility of deriving the tangent linear model and its adjoint. DALECv2 was designed to take into account this requirement, in particular by replacing the phenology process of the DALEC deciduous model in order to obtain differentiable processes. The model resolution matrix and the gradient of the cost function, including the additional term encoding the EDCs, are computed using adjoint techniques. Despite the increasing capacities offered by automatic differentiation tools, deriving and maintaining an adjoint code can be a complicated task, and, besides its limiting hypothesis, this is certainly one of the main reason for choosing alternatives to 4DVAR. In a forthcoming paper, we use ensem-

ble methods to approximate the gradient of the cost function and to derive approximate resolution matrices, and the experiments presented in this paper are reproduced. The approach, which no longer requires the adjoint, shows very promising results: firstly in terms of estimating parameters, and secondly in terms of computation time by using graphic processing units (GPUs) to perform massive parallel computations.

Designing a global-scale experiment involving a coupling between DALEC and a transport model has been considered but is still at an early stage. As presented in Bloom and Williams (2015), the EDCs were originally introduced to constrain unresolved parameters at the site level where, in the absence of any other information, only MODIS LAI observations were available. In theory there is no restriction to readily apply the same constraints at a global scale; however, their efficiency highly depends on the nature of the coupling between the ecosystem model and the transport model, and on the observation streams considered. Nonetheless in this context 4DVAR remains the only reasonable method to consider in terms of computer resources, and our study demonstrates that the current research efforts to develop regularisation strategies fit well into the variational framework.

## 8 Conclusions

We used DALECv2 and combined multiple data streams – MODIS monthly LAI and monthly NEE, GPP and RESP at an AmeriFlux site – together with ecological constraints to estimate model parameters and initial conditions and to provide uncertainty characterisation for predicted fluxes. DALECv2 is a simple model. It represents the basic processes at the heart of more sophisticated models of the carbon cycle, and, besides its large modelling skills, its simplicity allows for close mathematical scrutiny. Here we adopted a variational approach where the tangent linear model and its adjoint play a major role in (1) facilitating a linear analysis which allows one to understand the nature of the ill-posed problem and to evaluate strategies to regularise it and (2) finding a posterior distribution for the state variables.

We performed a sensitivity analysis using a direct method that consists of studying the first-order derivatives of the output computed using an adjoint method. A sensitivity analysis is a prerequisite to any work with a model, but there is a paucity of literature on this topic in connection with DALEC. Our analysis reveals generic issues that will be encountered in many inverse modelling strategies. Studying the first-order inverse problem, we discussed how noise affects the stability of the solution and we illustrated a simple regularisation method. We then introduced the notion of a model resolution matrix and showed how this can be used to diagnose the ill-posedness of an inverse problem and evaluate the result of regularisation strategies. While some of our findings may be anticipated in the framework of a simple model, it is important to describe these tools and their interpretation, as similar

analyses can be readily applied to a wide range of more complex models.

Bloom and Williams (2015) proved the advantage of the EDCs in constraining poorly resolved components of the carbon cycle and recommended their use for inverse modelling problems. We successfully incorporated the EDCs within the context of variational data assimilation. Our results confirm that the EDCs regularise an otherwise ill-posed problem and efficiently reduce the uncertainty of predicted fluxes. Moreover, our modification to DALECv2, DALEC-SP, which includes a spin-up period, offers an alternative to some EDCs that facilitates the variational approach.

This study did not aim at providing an exhaustive account on the capability of variational tools or exploring all aspects of the EDCs for the inverse problem for DALEC. The objectives were to use 4DVAR and show that it offers a suitable framework to solve efficiently, robustly and quickly the inverse problem for DALEC, and to present a methodology to analyse some issues that affect most methods based on Bayesian inference.

# References

Bloom, A. A. and Williams, M.: Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model-data fusion framework, Biogeosciences, 12, 1299–1315, https://doi.org/10.5194/bg-12-1299-2015, 2015.

Chuter, A. M., Aston, P. J., Skeldon, A. C., and Roulstone, I.: A dynamical systems analysis of the data assimilation linked ecosystem carbon (DALEC) models, Chaos, 25, 036401, https://doi.org/10.1063/1.4897912, 2015.

Fox, A., Williams, M., Richardson, A. D., Cameron, D., Gove, J. H., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C. M., and Van Wijk, M. T.: The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data, Agr. Forest Meteorol., 149, 1597–1615, https://doi.org/10.1016/j.agrformet.2009.05.002, 2009.

Giering, R. and Kaminski, T.: Recipes for Adjoint Code Construction, ACM Trans. Math. Softw., 24, 437–474, https://doi.org/10.1145/293686.293695, 1998.

Golub, G. H. and Van Loan, C. F.: Matrix Computations, 3rd Edn., Johns Hopkins University Press, Baltimore, MD, USA, 1996.

Hadamard, J.: Lectures on Cauchy's problem in linear partial differential equations, Yale University Press, Yale, 1923.

Hansen, P. C.: Regularization Tools version 4.0 for Matlab 7.3, Numer. Algorit., 46, 189–194, https://doi.org/10.1007/s11075-007-9136-9, 2007.

Hansen, P. C. and O'Leary, D. P.: The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems, SIAM J. Scient. Comput., 14, 1487–1503, https://doi.org/10.1137/0914086, 1993.

Hill, T. C., Ryan, E., and Williams, M.: The use of $CO_2$ flux time series for parameter and carbon stock estimation in carbon cycle research, Global Change Biol., 18, 179–193, https://doi.org/10.1111/j.1365-2486.2011.02511.x, 2012.

Johnson, C., Hoskins, B. J., and Nichols, N. K.: A singular vector perspective of 4D-Var: Filtering and interpolation, Q. J. Roy. Meteorol. Soc., 131, 1–19, https://doi.org/10.1256/qj.03.231, 2005.

Kemp, S., Scholze, M., Ziehn, T., and Kaminski, T.: Limiting the parameter space in the Carbon Cycle Data Assimilation System (CCDAS), Geosci. Model Dev., 7, 1609–1619, https://doi.org/10.5194/gmd-7-1609-2014, 2014.

MacBean, N., Peylin, P., Chevallier, F., Scholze, M., and Schürmann, G.: Consistent assimilation of multiple data streams in a carbon cycle data assimilation system, Geosci. Model Dev., 9, 3569–3588, https://doi.org/10.5194/gmd-9-3569-2016, 2016.

Menke, W.: Geophysical Data Analysis: Discrete Inverse Theory, Academic Press, New York, https://doi.org/10.1016/B978-0-12-490920-5.50001-6, 1984.

Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., and Widmann, H.: Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS), Global Biogeochem. Cyc., 19, gB2026, https://doi.org/10.1029/2004GB002254, 2005.

Richardson, A. D., Williams, M., Hollinger, D. Y., Moore, D. J. P., Dail, D. B., Davidson, E. A., Scott, N. A., Evans, R. S., Hughes, H., Lee, J. T., Rodrigues, C., and Savage, K.: Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints, Oecologia, 164, 25–40, https://doi.org/10.1007/s00442-010-1628-y, 2010.

Roese-Koerner, L., Devaraju, B., Sneeuw, N., and Schuh, W.-D.: A stochastic framework for inequality constrained estimation, J. Geodesy, 86, 1005–1018, https://doi.org/10.1007/s00190-012-0560-9, 2012.

Safta, C., Ricciuto, D. M., Sargsyan, K., Debusschere, B., Najm, H. N., Williams, M., and Thornton, P. E.: Global sensitivity analysis, probabilistic calibration, and predictive assessment for the data assimilation linked ecosystem carbon model, Geosci. Model Dev., 8, 1899–1918, https://doi.org/10.5194/gmd-8-1899-2015, 2015.

Williams, M., Rastetter, E. B., Fernandes, D. N., Goulden, M. L., Shaver, G. R., and Johnson, L. C.: Predicting gross primary productivity in terrestrial ecosystems, Ecol. Appl., 7, 882–894, https://doi.org/10.1890/1051-0761(1997)007[0882:PGPPIT]2.0.CO;2, 1997.

Williams, M., Schwarz, P., Law, B., Irvine, J., and Kurpius, M.: An improved analysis of forest carbon dynamics using data assimilation, Global Change Biol., 11, 89–105, https://doi.org/10.1111/j.1365-2486.2004.00891.x, 2005.

Zhu, Q. and Zhuang, Q.: Parameterization and sensitivity analysis of a process-based terrestrial ecosystem model using adjoint method, J. Adv. Model. Earth Syst., 6, 315–331, https://doi.org/10.1002/2013MS000241, 2014.

Ziehn, T., Scholze, M., and Knorr, W.: On the capability of Monte Carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models, Global Biogeochem. Cy., 26, gB3025, https://doi.org/10.1029/2011GB004185, 2012.