RMetS
Royal Meteorological Society

Improved analysis-error covariance matrix for high-dimensional variational inversions: application to source estimation using a 3D atmospheric transport model

N. Bousserez,[a]* D. K. Henze,[a] A. Perkins,[a] K. W. Bowman,[b] M. Lee,[b] J. Liu,[b] F. Deng[c] and D. B. A. Jones[c,d]

[a]*Department of Mechanical Engineering, University of Colorado, Boulder, USA*
[b]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA*
[c]*Department of Physics, University of Toronto, Ontario, Canada*
[d]*JIFRESSE, University of California, Los Angeles, USA*

*Correspondence to: N. Bousserez, ECES, 1111 Engineering Drive UCB 427, Boulder, CO 80309, USA.
E-mail: nicolas.bousserez@colorado.edu

Variational methods are widely used to solve geophysical inverse problems. Although gradient-based minimization algorithms are available for high-dimensional problems (dimension $> 10^6$), they do not provide an estimate of the errors in the optimal solution. In this study, we assess the performance of several numerical methods to approximate the analysis-error covariance matrix, assuming reasonably linear models. The evaluation is performed for a CO_2 flux estimation problem using synthetic remote-sensing observations of CO_2 columns. A low-dimensional experiment is considered in order to compare the analysis error approximations to a full-rank finite-difference inverse Hessian estimate, followed by a realistic high-dimensional application. Two stochastic approaches, a Monte-Carlo simulation and a method based on random gradients of the cost function, produced analysis error variances with a relative error $< 10\%$. The long-distance error correlations due to sampling noise are significantly less pronounced for the gradient-based randomization, which is also particularly attractive when implemented in parallel. Deterministic evaluations of the inverse Hessian using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm are also tested. While existing BFGS preconditioning techniques yield poor approximations of the error variances (relative error $> 120\%$), a new preconditioner that efficiently accumulates information on the diagonal of the inverse Hessian dramatically improves the results (relative error $< 50\%$). Furthermore, performing several cycles of the BFGS algorithm using the same gradient and vector pairs enhances its performance (relative error $< 30\%$) and is necessary to obtain convergence. Leveraging those findings, we proposed a BFGS hybrid approach which combines the new preconditioner with several BFGS cycles using information from a few (3–5) Monte-Carlo simulations. Its performance is comparable to the stochastic approximations for the low-dimensional case, while good scalability is obtained for the high-dimensional experiment. Potential applications of these new BFGS methods range from characterizing the information content of high-dimensional inverse problems to improving the convergence rate of current minimization algorithms.

Key Words: data assimilation; variational methods; uncertainty quantification; analysis error; BFGS algorithm

Received 26 November 2013; Revised 3 November 2014; Accepted 11 November 2014; Published online in Wiley Online Library

1. Introduction

Many inverse problems in geophysics are solved within the classical Bayesian framework, in which a maximum likelihood estimate is derived from a combination of observational and prior information as well as a model describing the physics relating the two quantities (Enting, 2002; Tarantola, 2005). Under the assumptions that the model is reasonably linear and that the observational and prior probability distribution functions (pdfs) are Gaussian (and unbiased with respect to the true state), the posterior pdfs are also Gaussian and the solution to the inverse problem is obtained by minimizing the cost

function

$$J(\mathbf{x}) = \frac{1}{2}(H\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{R}^{-1}(H\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_{\mathrm{b}})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_{\mathrm{b}}), \quad (1)$$

where $\mathbf{x}_{\mathrm{b}}$ is the prior vector, defined in the control space E of dimension n, $\mathbf{x}$ belongs to E, $\mathbf{y}$ is the observation vector, defined on the observations vector space F of dimension p, $H : E \rightarrow F$ is the forward model operator (also called observational operator), which associates to any vector in E its corresponding observation in F, and $\mathbf{R}$ and $\mathbf{B}$ are the covariance matrices of observation and prior errors with dimension $(p \times p)$ and $(n \times n)$ respectively.

The argument of the minimum of Eq. (1) is called the analysis and is referred to as $\mathbf{x}_{\mathrm{a}}$. For high-dimensional systems, such as those encountered in numerical weather forecast (NWF) applications ($n > 10^6$), $\mathbf{x}_{\mathrm{a}}$ is often estimated using iterative gradient minimization algorithms that require at each iteration the calculation of the gradient of J with respect to $\mathbf{x}$:

$$\nabla J(\mathbf{x}) = \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(H\mathbf{x}-\mathbf{y}) + \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_{\mathrm{b}}), \quad (2)$$

where $\mathbf{H}$ is the tangent-linear of the forward model, and $\mathbf{H}^{\mathrm{T}}$ its adjoint.

For high-dimensional problems, $\mathbf{H}$, $\mathbf{B}$ and $\mathbf{R}$ cannot be represented explicitly and are usually decomposed into lower-dimension operators (Bannister, 2008; Singh *et al.*, 2011). Adjoint techniques allow efficient numerical computations of Eq. (2) by decomposing $\mathbf{H}^{\mathrm{T}}$ into a chain of lower-dimension adjoint operators integrated backward in time (Henze *et al.*, 2007). Second-order information is also crucial in data assimilation. It allows for computation of the analysis and forecast errors, as well as the sensitivity of the inversion to input parameters (e.g. observations, error statistics) and to improve the convergence rate of the minimization (Courtier *et al.*, 1994; Le Dimet *et al.*, 2002; Le Dimet and Shutyaev, 2005; Daescu, 2008). Several approaches exist to derive second-order information of large systems, including second-order adjoint, randomization estimates, quasi-Newton methods, and singular value analysis based on the Hessian-vector product (Davidon, 1991; Le Dimet *et al.*, 2002). These methods differ in term of their computational efficiency, their accuracy, and their development requirements. In variational methods, where the analysis-error covariance is not directly obtained, being able to generate a good approximation of this matrix is still a major challenge for high-dimensional problems. Assuming that the forward model operator H can be approximated by a linear operator in a neighbourhood of the analysis $\mathbf{x}_{\mathrm{a}}$, the analysis-error covariance matrix $\mathbf{P}^{\mathrm{a}}$ is equal to the inverse Hessian of the cost function at $\mathbf{x}_{\mathrm{a}}$ (Eq. (1)) (Tarantola, 2005):

$$\mathbf{P}^{\mathrm{a}} = (\nabla^2 J)^{-1}(\mathbf{x}_{\mathrm{a}}) = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}, \quad (3)$$

where $\mathbf{P}^{a}$ has dimension $n \times n$. Estimating each element of $\mathbf{P}^{\mathrm{a}}$ using finite-difference approximations would require $n(n+3)/2+1$ forward model integrations (Nocedal and Wright, 2006). With an adjoint model, the cost of the calculation reduces to $n+1$ forward model and $n+1$ adjoint model integrations, which is still prohibitive in most cases. Therefore, $\mathbf{P}^{\mathrm{a}}$ is usually approximated using low-rank estimates of the inverse Hessian (Fisher and Courtier, 1995; Müller and Stavrakou, 2005).

When the model is a linear operator ($H = \mathbf{H}$), the cost function (Eq. (1)) is perfectly quadratic and the conjugate-gradient (CG) algorithm is usually considered the best approach for the minimization of Eq. (1) (Nocedal and Wright, 2006). This method is closely related to the Lanczos algorithm and can provide the leading eigenvalues and eigenvectors of the Hessian as a by-product of the minimization. This allows an efficient approximation to the analysis-error covariance matrix (Fisher and Courtier, 1995). In the context of cyclic data assimilation, this Lanczos eigendecomposition can also be used to precondition the CG algorithm for subsequent assimilation windows and therefore improve the rate of convergence of the minimization.

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm was developed in the context of non-quadratic quasi-Newton minimization methods (Broyden, 1969). The method extracts local curvature information from a sequence of gradient and vector pairs to calculate the line-search direction at each iteration. For quadratic problems and exact line searches, it can be shown that the BFGS and conjugate gradient algorithms are equivalent, in the sense that they produce the same search direction at each iteration (Nazareth, 1979). Its efficiency and also a number of desirable mathematical properties make this algorithm one of the best to approximate the inverse Hessian (Dennis and More, 1977). Since each iteration consists of a rank-two update, and the number of available gradient and vector pairs is generally small compared to the dimension of the problem, the rank of the approximated inverse Hessian is usually very low ($\ll n$). However, significant improvements to the BFGS inverse Hessian estimate may be achieved by choosing a suitable starting matrix in the algorithm, a technique referred as 'preconditioning' (Gilbert and Lemaréchal, 1989). Preconditioning is especially critical in the context of limited-memory BFGS (L-BFGS) optimization, when only a few gradient and vector pairs can be stored in computer core-memory (Veersé *et al.*, 2000). However, it can also have a significant impact on the inverse Hessian estimate in the full-memory case, which is investigated in this study.

The efficiency of different preconditioning methods is highly problem-dependent. A classical approach is to use an initial identity matrix scaled by the so-called Oren–Spedicato scalar (Nocedal and Wright, 2006). More sophisticated diagonal preconditioners have been proposed in the literature, allowing significant improvements in the minimization or inverse Hessian approximation performances (Gilbert and Lemaréchal, 1989; Veersé *et al.*, 2000; Leong and Chen, 2013). Recent studies have also proposed to combine information from several minimizations. For instance, in the context of incremental 4D-Var data assimilation, Tshimanga *et al.* (2008) and Gratton *et al.* (2011b) exploit information gained from one CG minimization to precondition the CG minimization of the next outer iteration. They use a class of Limited-Memory Preconditioners (LMPs), of which the quasi-Newton BFGS update is a particular member.

Besides deterministic algorithms such as BFGS, other approaches for approximating the covariance matrix of analysis errors make use of the probabilistic nature of the Bayesian inverse problem. Recently, the stochastic method employing direct Monte-Carlo calculation was used to estimate analysis errors in a trace-gas variational inversion (Chevallier *et al.*, 2007). In practice, reliable estimates require at least 50 perturbed inversions to be performed, which renders this method computationally expensive and may preclude its use for time-limited applications (e.g. operational NWP). Another study by Desroziers *et al.* (2005) applied a randomization method to calculate the total error reduction associated with a subset of observations. In practice, this method allows estimation of only the trace of the inverse Hessian and not the individual diagonal elements (error variances). Rabier and Courtier (1992) proposed another approach based on an ensemble of perturbed gradients that allows one to approximate the observational term ($\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$) of the inverse Hessian expression (3) (Courtier *et al.*, 1994; Fisher and Courtier, 1995). One drawback of stochastic approaches is that they produce spurious long-range error correlations, a well-known problem in sequential data assimilation methods such as the Ensemble Kalman Filter (Anderson, 2001; Evensen, 2007). The issue of sampling noise can be addressed by methods such as localization by Schur products (Houtekamer and Mitchell, 2001) or the wavelet-diagonal approach (Buehner, 2012).

Finally, a recent study by Gejadze *et al.* (2011) combined stochastic and deterministic methods to estimate the analysis error covariance in a variational setting. One implementation of

their new 'effective inverse Hessian' (EIH) approach consists of computing a sample mean of the inverse Hessian for different perturbations of the true state. They demonstrate that their method reduces both the linearization error obtained when using the inverse Hessian and the sampling error associated with stochastic estimates.

Here we present an evaluation of stochastic and deterministic approaches to estimate the covariance matrix of analysis errors in a large-scale variational inverse problem. For the stochastic methods, a classical Monte-Carlo calculation as well as an approach based on random gradients of the 4D-Var cost function are considered, while the deterministic approach is based on BFGS approximations of the inverse Hessian of the functional. First, we consider a low-dimensional inverse problem in atmospheric transport so that a full-rank estimate of the inverse Hessian using finite differences is computationally feasible. This approximation is used as a reference to evaluate the efficiency of the different algorithms considered. Leveraging results obtained for existing algorithms, we propose a new diagonal BFGS preconditioner as well as simple techniques to dramatically improve the BFGS inverse Hessian estimate. The computational requirements and efficiency of each method are discussed. We also test the scalability of the proposed algorithms by applying them to a high-dimensional CO_2 flux inversion system, the Carbon Monitoring System (CMS) Flux Pilot Project, which is currently under development by the National Aeronautics and Space Administration (NASA; Liu *et al.*, 2014)

The article is structured as follows. In section 2, we present two existing stochastic approaches to estimate the covariance matrix of analysis errors. Section 3 describes the BFGS algorithm and some of its fundamental mathematical properties. It also reviews commonly used preconditioning methods and proposes a new preconditioning technique. Results are presented in section 4 for both the low-dimensional test case and the realistic high-dimensional experiment. Conclusions are provided in section 5.

2. Stochastic approximations

2.1. Monte-Carlo method

One way to estimate the analysis-error covariance matrix ($\mathbf{P}^{\mathrm{a}}$) is by directly calculating a sample estimate:

$$\begin{aligned}\mathbf{P}^{\mathrm{a}} &= \overline{(\mathbf{x}_{\mathrm{a}}-\mathbf{x}_{\mathrm{t}})(\mathbf{x}_{\mathrm{a}}-\mathbf{x}_{\mathrm{t}})^{\mathrm{T}}} &(4)\\ &\approx \overline{(\mathbf{x}'_{\mathrm{a}}-\mathbf{x}_{\mathrm{ref}})(\mathbf{x}'_{\mathrm{a}}-\mathbf{x}_{\mathrm{ref}})^{\mathrm{T}}} &(5)\\ &\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_{\mathrm{a}}^{i\prime}-\mathbf{x}_{\mathrm{ref}})(\mathbf{x}_{\mathrm{a}}^{i\prime}-\mathbf{x}_{\mathrm{ref}})^{\mathrm{T}},\end{aligned}$$

where $\overline{x}$ is the expectation of x, $\mathbf{x}_{\mathrm{t}}$ is the true state, $\mathbf{x}_{\mathrm{ref}}$ is a known reference state, and each analysis $\mathbf{x}_{\mathrm{a}}^{i\prime}$ is obtained by perturbing the reference state $\mathbf{x}_{\mathrm{ref}}$ and the associated observations ($\mathbf{y}_{\mathrm{ref}}=H(\mathbf{x}_{\mathrm{ref}})$) according to the assumed error statistics for the prior ($\mathbf{B}$) and the observations ($\mathbf{R}$), respectively (Chevallier *et al.*, 2007). Formally, it can be written as

$$\left.\begin{aligned}\mathbf{y}^{i} &= \mathbf{y}_{\mathrm{ref}}+\mathbf{V}^{\mathrm{T}}\mathbf{v}^{1/2}\mathbf{p}^{i},\\ \mathbf{x}_{\mathrm{b}}^{i} &= \mathbf{x}_{\mathrm{ref}}+\mathbf{W}^{\mathrm{T}}\mathbf{w}^{1/2}\mathbf{q}^{i},\end{aligned}\right\} \quad (6)$$

where $\mathbf{V}$ and $\mathbf{v}$ are the eigenvector and eigenvalue matrices of $\mathbf{R}$, respectively, $\mathbf{W}$ and $\mathbf{w}$ are the eigenvector and eigenvalue matrices of $\mathbf{B}$, respectively, and i denotes a realization of the random variables $\mathbf{p}$ and $\mathbf{q}$ which both follow a standard normal distribution.

An important implicit assumption is made when using this algorithm to estimate $\mathbf{P}^{\mathrm{a}}$. While the reference state $\mathbf{x}_{\mathrm{ref}}$ being used is different from the true state $\mathbf{x}_{\mathrm{t}}$, it is assumed that $\mathbf{P}^{\mathrm{a}}$ does not depend on $\mathbf{x}_{\mathrm{ref}}$, at least in the vicinity of $\mathbf{x}_{\mathrm{t}}$. In practice this is equivalent to assuming that the model H is linear in the vicinity of $\mathbf{x}_{\mathrm{t}}$. Typical choices for $\mathbf{x}_{\mathrm{ref}}$ are the prior $\mathbf{x}_{\mathrm{b}}$ or the analysis $\mathbf{x}_{\mathrm{a}}$, which represent the best estimates of the true state $\mathbf{x}_{\mathrm{t}}$ available before and after the inversion, respectively.

A consequence of the Gaussian distributions assumed for the prior and the observations is that each element $(\mathbf{x}_{\mathrm{a}}^{i}-\mathbf{x}_{\mathrm{t}})_j$ of the analysis error itself follows a Gaussian distribution $\mathcal{N}(0,\sigma_{\mathrm{a}}^{j})$, where σ_{a}^{j} represents the standard analysis error of element j. Therefore the sum

$$\sum_{i=1}^{N}\frac{(\mathbf{x}_{\mathrm{a}}^{i}-\mathbf{x}_{\mathrm{t}})_j^2}{(\sigma_{\mathrm{a}}^{j})^2}$$

follows a chi-squared distribution with N degrees of freedom. Basic statistics show that the relative standard error in the estimated variance (σ_{a}^2) and standard deviation (σ_{a}) are $\sqrt{2/N}$ and $1/\sqrt{2N}$, respectively. This shows, for example, that a 10% relative standard error for σ_{a} can be obtained with an ensemble of only 50 vectors. Notably, the relative standard error in σ_{a} is independent of the dimension (n) of the problem. However, approximating the error covariances from a sample estimate is a more difficult task than approximating the variance terms (e.g. Berre and Desroziers, 2010).

2.2. Gradient-based randomization

Rabier and Courtier (1992) proposed a stochastic approach to estimate the observational term $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ in Eq. (3). It is based on the equivalence between the observational term of the Hessian matrix and the covariance matrix of gradients of the cost function, and can be derived as follows. Consider

$$\nabla J(\mathbf{x}_{\mathrm{b}}) = \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\xi}, \quad (7)$$

where $\xi \equiv H(\mathbf{x}_{\mathrm{b}})-y$. We can generate an ensemble of realizations of y (pseudo-observations) such that

$$\left.\begin{aligned}\overline{\boldsymbol{\xi}} &= \mathbf{0},\\ \overline{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}} &= \mathbf{R}.\end{aligned}\right\} \quad (8)$$

We obtain

$$\overline{\nabla J(\mathbf{x}_{\mathrm{b}})\nabla J(\mathbf{x}_{\mathrm{b}})^{\mathrm{T}}} = \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\overline{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}}\mathbf{R}^{-1}\mathbf{H}. \quad (9)$$

Therefore we can write

$$\begin{aligned}\nabla^2 J(\mathbf{x}_a) &= \mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\\ &= \mathbf{B}^{-1}+\overline{\nabla J(\mathbf{x}_{\mathrm{b}})\nabla J(\mathbf{x}_{\mathrm{b}})^{\mathrm{T}}}. \quad (10)\end{aligned}$$

The covariance matrix of gradients in the right-hand side can be approximated using a sample of gradients. The inverse Hessian $(\nabla^2 J)^{-1}(\mathbf{x}_{\mathrm{a}})$ can then be obtained by inverting the above expression using (see Appendix) the Shermann–Morrison–Woodbury formula (Sherman and Morrison, 1949). For practical implementations of the above method when $\mathbf{R}$ is not diagonal, we may use the following change of variable:

$$\overline{\nabla J(\mathbf{x}_{\mathrm{b}})\nabla J(\mathbf{x}_{\mathrm{b}})^{\mathrm{T}}} = \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{v}^{1/2}\overline{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}}\mathbf{v}^{1/2}\mathbf{V}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}, \quad (11)$$

where $\mathbf{V}$ and $\mathbf{v}$ are the eigenvector and eigenvalue matrices of $\mathbf{R}$, respectively, and $\boldsymbol{\varepsilon}$ is a vector of independent random numbers with standard normal distribution.

Each element $\{\nabla J(\mathbf{x}_{\mathrm{b}})\}_j = (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\xi})_j$ of the random variable $\nabla J(\mathbf{x}_{\mathrm{b}})$ is a linear combination of elements of $\boldsymbol{\xi}$, which follow a Gaussian distribution. Therefore $\{\nabla J(\mathbf{x}_{\mathrm{b}})\}_j$ is itself a Gaussian random variable. Thus the standard error for the sample estimate of the variance

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\xi}^{i})_j^2 \qquad \text{is} \qquad (\sigma_j')^2\sqrt{\frac{2}{N}},$$

where $(\sigma_j')^2$ represents the variance of $\{\nabla J(\mathbf{x}_b)\}_j$. However, since we seek to estimate the inverse of the stochastically approximated Hessian matrix, the resulting estimator of the inverse Hessian may be far from a chi-square variable. Therefore, as opposed to the Monte-Carlo case, the sampled estimate of the diagonal of the inverse Hessian may be significantly biased and the number of samples required to reach a given relative standard error depends upon the property of the problem (section 4.1.2). Finally, since for high-dimensional problems the Hessian matrix cannot be explicitly stored in memory nor directly inverted, efficient numerical algorithms to extract information from this matrix and perform algebraic operations are proposed in the Appendix.

3. The BFGS algorithm

3.1. Principle

The BFGS algorithm (Broyden, 1969) seeks to approximate the inverse Hessian matrix of the cost function using gradient and vector pairs generated during the minimization. To be consistent with the notations commonly used in the literature while ensuring clarity of the present article, in the following the inverse Hessian will be noted as $\widehat{\mathbf{H}}$, and needs to be distinguished from the forward model operator H defined in previous sections. Likewise, the vectors $\mathbf{y}$ with a subscript represent a difference of gradients (see below), and not the vector of observations defined above. The inverse Hessian update at each iteration k is given by (Nocedal and Wright, 2006)

$$\widehat{\mathbf{H}}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^{\mathrm{T}})\widehat{\mathbf{H}}_k(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}}) + \rho_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}}, \tag{12}$$

with

$$\rho_k = \frac{1}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k},$$

and

$$\begin{aligned} \mathbf{s}_k &= \mathbf{x}_{k+1} - \mathbf{x}_k\,, \\ \mathbf{y}_k &= \nabla J(\mathbf{x}_{k+1}) - \nabla J(\mathbf{x}_k)\,, \end{aligned}$$

where J is the cost function of the minimization problem.

At each iteration, the minimization update is given by

$$\left.\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{p}_k\,, \\ \mathbf{p}_k &= \widehat{\mathbf{H}}_k \nabla J(\mathbf{x}_k), \end{aligned}\right\} \tag{13}$$

where α_k is chosen to satisfy the so-called Wolfe conditions:

$$\left.\begin{aligned} J(\mathbf{x}_k + \alpha_k \mathbf{p}_k) &\le J(\mathbf{x}_k) + c_1 \alpha_k \nabla J(\mathbf{x}_k)^{\mathrm{T}} \mathbf{p}_k, \\ \nabla J(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^{\mathrm{T}} \mathbf{p}_k &\ge c_2 \nabla J(\mathbf{x}_k) \mathbf{p}_k, \end{aligned}\right\} \tag{14}$$

with $0 < c_1 < c_2 < 1$.

For a general non-quadratic cost function, superlinear convergence of the BFGS method can be demonstrated (Dennis and More, 1977). Conditions for convergence of the inverse Hessian approximation are more restrictive. If $\mathbf{y}_k$ and $\mathbf{s}_k$ are generated using a quasi-Newton algorithm with exact line searches, then the BFGS approximation to the inverse Hessian converges to the true inverse Hessian in at most n iterations (Bonnans *et al.*, 2006). In the nonlinear case, exact line searches are impossible, and therefore there is no guarantee that convergence to the inverse Hessian can be reached in a finite number of steps.

3.2. BFGS preconditioning

Note that the initial inverse Hessian $\widehat{\mathbf{H}}_0$ in Eq. (12) can be freely chosen. Moreover, at each iteration k, a new initial matrix, noted $\widehat{\mathbf{H}}_k^0$, can be used. The choice of the initial inverse Hessian plays an important role in the performance of the BFGS algorithm. Preconditioning methods have been mainly developed in the context of limited-memory BFGS, where computer memory capacity allows only a small number of gradient and vector pairs to be stored. Assuming a maximum of m gradient and vector pairs are available at each iteration k, the following formula can be used to update $\widehat{\mathbf{H}}_k$:

$$\begin{aligned} \widehat{\mathbf{H}}_k = {} & (\mathbf{V}_{k-1}^{\mathrm{T}} \ldots \mathbf{V}_{k-m}^{\mathrm{T}}) \widehat{\mathbf{H}}_k^0 (\mathbf{V}_{k-m} \ldots \mathbf{V}_{k-1}) \\ & + \rho_{k-m} (\mathbf{V}_{k-1}^{\mathrm{T}} \ldots \mathbf{V}_{k-m+1}^{\mathrm{T}}) \mathbf{s}_{k-m} \mathbf{s}_{k-m}^{\mathrm{T}} (\mathbf{V}_{k-m+1} \ldots \mathbf{V}_{k-1}) \\ & + \rho_{k-m+1} (\mathbf{V}_{k-1}^{\mathrm{T}} \ldots \mathbf{V}_{k-m+2}^{\mathrm{T}}) \mathbf{s}_{k-m+1} \mathbf{s}_{k-m+1}^{\mathrm{T}} (\mathbf{V}_{k-m+2} \ldots \mathbf{V}_{k-1}) \\ & + \ldots \\ & + \rho_{k-1} \mathbf{s}_{k-1} \mathbf{s}_{k-1}^{\mathrm{T}}\,, \end{aligned} \tag{15}$$

with $\mathbf{V}_k = I - \rho_k \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}}$. One advantage of this formulation is that it allows an efficient two-loop recursion algorithm to compute any matrix–vector product $\widehat{\mathbf{H}}_k \mathbf{v}$. This type of formula is referred to as 'implicit' because the inverse Hessian matrix is never explicitly formed, but rather implicitly known through the ensemble of gradient and vector pairs $(\mathbf{y}_k, \mathbf{s}_k)$, which can be used to extract any element of the matrix or to perform matrix–vector product (see Appendix).

Finding the appropriate method to define $\widehat{\mathbf{H}}_k^0$ (the preconditioner) is critical to obtain good rates of convergence for the minimization. A general principle is that $\widehat{\mathbf{H}}_k^0$ should retain as much information as possible from previous BFGS iterations. It is often set to some multiple of the identity, $\gamma \mathbf{I}$, although more sophisticated diagonal preconditioners have also been developed (see below). The choice of the preconditioner is highly problem-dependent, and in this section we present some of the commonly used preconditioning techniques.

A method that has proved effective in practice is to use the so-called Oren–Spedicato scalar γ_k:

$$\left.\begin{aligned} \gamma_k &= \frac{\mathbf{s}_{k-1}^{\mathrm{T}} \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^{\mathrm{T}} \mathbf{y}_{k-1}}, \\ \widehat{\mathbf{H}}_k^0 &= \gamma_k \mathbf{I}\,. \end{aligned}\right\} \tag{16}$$

In addition to its good performance, this preconditioner has some desirable mathematical properties, among which is the fact that γ_k is the Rayleigh quotient of $\overline{\overline{\mathbf{H}}}$ in the direction $\mathbf{y}_k$, i.e. $\gamma_k = \mathbf{y}_{k-1}^{\mathrm{T}} \widehat{\mathbf{H}} \mathbf{y}_{k-1} / \mathbf{y}_{k-1}^{\mathrm{T}} \mathbf{y}_{k-1}$, where

$$\overline{\overline{\mathbf{H}}} \equiv \left(\int_0^1 \nabla^2 J(\mathbf{x}_{k-1} + t\, \mathbf{s}_{k-1})\, \mathrm{d}t \right)^{-1}. \tag{17}$$

The reader is referred to Gilbert and Lemaréchal (1989) for more details about the useful properties of γ_k. In the following we will refer to this preconditioning method as INIT_O.

It has been recently shown (Leong and Chen, 2013) that the Oren–Spedicato preconditioner belongs to a broader class of diagonal preconditioners of the form

$$\widehat{\mathbf{H}}_k^0 = \alpha_k \mathbf{I} + \frac{(\mathbf{y}_{k-1}^{\mathrm{T}} \mathbf{s}_{k-1} - \alpha_k \mathbf{y}_{k-1}^{\mathrm{T}} \mathbf{y}_{k-1})}{\mathrm{tr}(\mathbf{Y}_{k-1}^2)} \mathbf{Y}_{k-1}, \tag{18}$$

where $\mathbf{Y}_{k-1} = \mathrm{diag}(\mathbf{y}_{k-1,1}^2, \ldots, \mathbf{y}_{k-1,n}^2)$, $\mathbf{y}_{k-1,i}$ is the ith component of the vector $\mathbf{y}_{k-1}$, and α_k is some positive scalar. This class is obtained using a variational technique based upon the weak-quasi-Newton equation of Dennis and Wolkowicz (1993):

$$\mathbf{y}_{k-1}^{\mathrm{T}} \widehat{\mathbf{H}}_k^0 \mathbf{y}_{k-1} = \mathbf{y}_{k-1}^{\mathrm{T}} \mathbf{s}_{k-1}. \tag{19}$$

Note that $\widehat{\mathbf{H}}_k^0 = \gamma_k \mathbf{I}$ correspond to $\alpha_k = \gamma_k$ in Eq. (18). In this study we will consider the particular case where $\alpha_k = 0$, and will refer to this preconditioner as LC_DIAG.

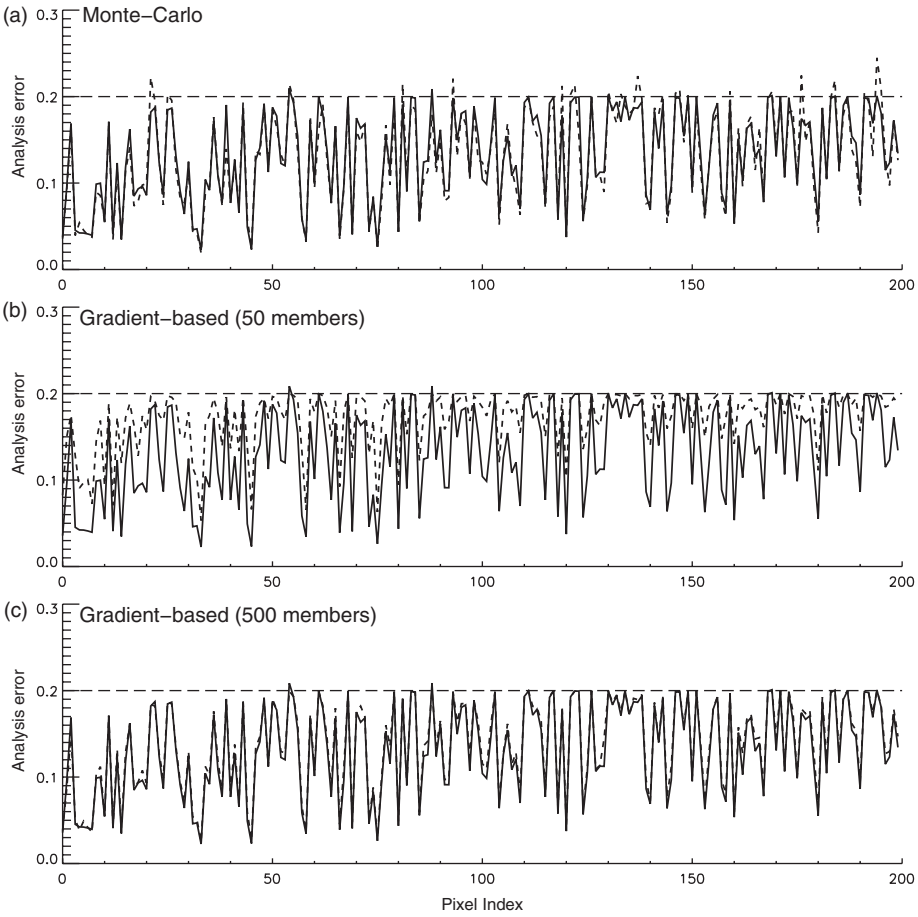

Figure 1. Analysis standard error estimates for all 200 optimized scaling factors using finite-difference calculation (solid line) and the stochastic methods (dashed line): (a) Monte-Carlo estimate using 50 members, (b) gradient-based estimate using 50 members, (c) gradient-based estimate using 500 members.

In addition to diagonal preconditioners based on the weak-quasi-Newton equation, other methods have been proposed by Gilbert and Lemaréchal (1989) and Veersé *et al.* (2000), where an initial diagonal matrix is updated following a diagonal version of the BFGS formula. Starting from $\widehat{\mathbf{H}}_1^0 = \gamma_0 \mathbf{I}$, the update is

$$\mathbf{D}_{k+1}^i = \mathbf{D}_k^i + \left(\frac{1}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k} + \frac{\mathbf{y}_k^{\mathrm{T}}\mathbf{D}_k\mathbf{y}_k}{(\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k)^2}\right)(\mathbf{s}_k^i)^2 - \frac{2\mathbf{D}_k^i\mathbf{y}_k^i\mathbf{s}_k^i}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k}, \qquad (20)$$

where $\mathbf{D}_k^i$ denotes the ith element of the diagonal preconditioner $\widehat{\mathbf{H}}_k^0$ at iteration k. Note that, although this update preserves positive-definiteness, it does not ensure that the quasi-Newton or weak-quasi-Newton equations are satisfied. In the following we will refer to this preconditioning method as GL_DIAG.

In this study we also propose a new diagonal preconditioner, which is more adapted to the problem of estimating the diagonal elements of the inverse Hessian matrix. At each iteration, the BFGS algorithm is restarted using a diagonal matrix made of the diagonal elements of the inverse Hessian estimate at the previous iteration. Formally, the diagonal preconditioner is obtained as:

$$\widehat{\mathbf{H}}_k^0 = \mathrm{diag}(\widehat{\mathbf{H}}_k). \qquad (21)$$

In the following we will refer to this preconditioner as LAST_DIAG. Note that more information from previous iterations is used to construct this preconditioner than GL_DIAG. While in GL_DIAG only the last $(\mathbf{y}, \mathbf{s})$ pair is used to update the preconditioner, in LAST_DIAG one needs to use all $(\mathbf{y}_{k-m+i}, \mathbf{s}_{k-m+i})_{(i=1,..,m)}$ pairs to update the diagonal elements of the inverse Hessian matrix at each iteration.

Finally, for high-dimensional problems it might not be possible to store explicitly the entire inverse Hessian matrix in computer memory. However, matrix–vector products and specific elements of this matrix can still be computed from the gradient and vector pairs $(\mathbf{y}_k, \mathbf{s}_k)$ using implicit formulations for the BFGS update. Some of these algorithms are presented in the Appendix.

Table 1. Performance statistics of the Monte-Carlo and gradient-based stochastic approaches with respect to the finite-difference inverse Hessian estimate.

Stochastic method	PCC	a, b	SDRE
Monte-Carlo (50 members)	0.96	1.00, −0.00	0.10
Gradient-based (50 members)	0.88	0.54, 0.10	0.70
Gradient-based (500 members)	0.99	0.95, 0.01	0.08

PCC= Pearson correlation coefficient.
a, b are coefficients in the linear fit $y = ax + b$.
SDRE= Standard deviation of relative error.

4. Numerical experiments

4.1. CO_2 flux inversion using remote-sensing pseudo-observations

In order to evaluate the performance of different methods to approximate the covariance matrix of analysis errors, we considered an inverse problem in atmospheric constituent transport. We conducted an Observing System Simulation Experiment (OSSE), where global fossil fuel CO_2 fluxes were estimated for the month of July 2009 using the GEOS-Chem global chemistry transport model (CTM) (http://www.geos-chem.org; accessed 20 November 2014) together with remote-sensing pseudo-observations of CO_2 columns from the Greenhouse gases

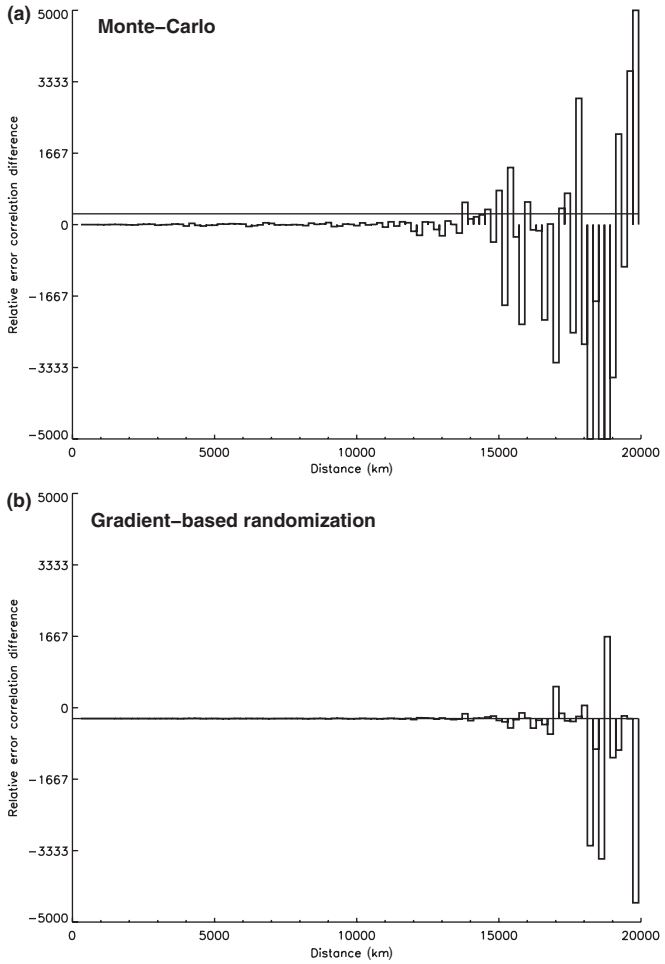

Figure 2. Relative differences between the analysis-error correlations calculated using finite-difference and the stochastic methods as a function of distance (values have been averaged over all scaling factors for each distance bin): (a) Monte-Carlo estimate, (b) gradient-based estimate.

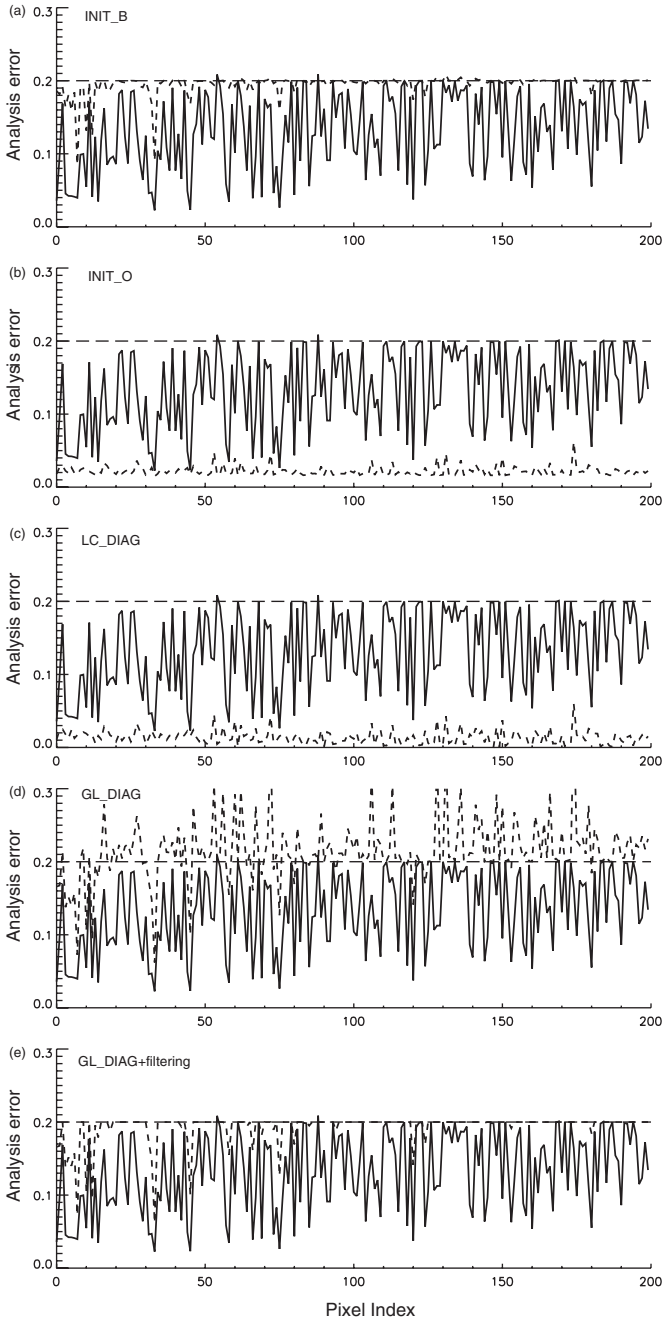

Figure 3. Analysis standard error estimates for all 200 optimized scaling factors using the finite-difference calculation (solid line) and different existing BFGS algorithms (dashed line): (a) INIT_B, (b) INIT_O, (c) LC_DIAG, (d) GL_DIAG with prior initialization, and (e) GL_DIAG with prior initialization and filtering. All BFGS calculations use the same 26 vector and gradient pairs obtained from one specific inversion.

Observing SATellite (GOSAT) instrument. In this inversion we optimized scaling factors applied to prior CO_2 fluxes instead of the fluxes themselves.

The GEOS-Chem CO_2 simulation has been described in recent studies (Nassar *et al.*, 2010, 2011; Liu *et al.*, 2014; Deng *et al.*, 2014). In our experiment we consider a $4^\circ \times 5^\circ$ horizontal model grid resolution. Meteorological fields are obtained from the GEOS (Goddard Earth Observing System) assimilated meteorology of the NASA Global Modeling Assimilation Office (GMAO). A more detailed description of the configuration of the GEOS-Chem CO_2 simulation and of the CO_2 data assimilation system can be found in Liu *et al.* (2014).

We use the Orbiting Carbon Observatory (ACOS) GOSAT CO_2 b2.9 retrievals (O'Dell *et al.*, 2012), which are based on measurements from the Thermal And Near-infrared Sensor for carbon Observation–Fourier Transform Spectrometer (TANSO-FTS; Yokota *et al.*, 2009). It follows a polar sun-synchronous orbit with a crossing at the Equator at 1300 local time, repeated every 3 days. The CO_2 retrievals consists of column-integrated dry mole fractions (X_{CO_2}). A more detailed description of the GOSAT retrievals can be found in O'Dell *et al.* (2012).

Here the GEOS-Chem model is used as the forward model that relates the CO_2 fluxes to the CO_2 concentration profiles, to which a GOSAT observational operator is applied to reproduce the observed CO_2 columns. The observational operator is obtained using the averaging kernels (**A**) of the retrieval. The standard error for the prior fluxes is set to 20% for all locations and we assume no error correlations (i.e. **B** is a diagonal matrix with 0.04 on the diagonal). Note that this assumption is an oversimplification. For instance, in the context of global biospheric CO_2 flux inversions, Chevallier *et al.* (2012) showed that, although spatial

Table 2. Performance statistics of several existing BFGS algorithms with respect to the finite-difference inverse Hessian estimate.

BFGS method (initialization)	PCC	*a*, *b*	SDRE
INIT_O (Eq. (16))	−0.07	0.00, 0.02	0.79
INIT_B	0.53	0.14, 0.17	1.39
GL_DIAG (Eq. (20)) (with filtering)	0.53	0.21, 0.16	1.22
LC_DIAG (Eq. (18))	−0.23	−0.04, 0.01	0.86

Column headings are as in Table 1.

error correlations are not significant, temporal error correlations from one month to the next can be as large as 0.6 and should therefore be included in the prior error covariance matrix **B**. The error statistics for the observations (**R**) are those provided by the GOSAT b2.9 retrievals, and it is assumed there are no correlations between observation errors (i.e. **R** is diagonal).

The short inversion window (1 month) for this experiment was chosen in order to lower the computational cost of the simulations, but this causes the inversion problem to be significantly ill-posed. To circumvent this difficulty, we artificially increased the sensitivity of the retrieval to the CO_2 surface concentrations (0–200 m) by a factor of 100. While this modification has no implication for the conclusions presented here, one must keep in mind that the results of this inversion have no physical significance.

The OSSE experiment consists of the following steps:

(i) Generate a set of GOSAT pseudo-observations for July 2009 from known global CO_2 fluxes. The vector of pseudo-observations is computed as:

$$y_{\mathrm{t}}^{i} = H^{i}\mathbf{x}_{\mathrm{t}} = \mathbf{A}^{i}(M^{i}\mathbf{x}_{\mathrm{t}} - \mathbf{c}_{\mathrm{a}}^{i}) + y_{\mathrm{a}}^{i}\,, \tag{22}$$

where H^i is the observational operator associated with the ith observed CO_2 column, $M^i\mathbf{x}_{\mathrm{t}}$ is the GEOS-Chem modelled CO_2 profile associated with that column generated using a vector $\mathbf{x}_{\mathrm{t}}$ of known CO_2 flux scaling factors, y_{a}^{i} and $\mathbf{c}_{\mathrm{a}}^{i}$ are the GOSAT *a priori* CO_2 column and *a priori* profile, respectively, and $\mathbf{A}^i$ is the GOSAT averaging kernel.

(ii) Apply a Gaussian random perturbation to each pseudo-observation, consistent with the error statistics defined by the covariance matrix of observation errors **R**:

$$y^{i} = y_{\mathrm{t}}^{i} + p^{i}, \tag{23}$$

where p^i is Gaussian noise with same standard deviation as the observation errors.

(iii) Apply a Gaussian random perturbation to each CO_2 flux scaling factor, consistent with the error statistics defined by the covariance matrix of background errors **B**:

$$\mathbf{x}_{\mathrm{b}} = \mathbf{x}_{\mathrm{t}} + \mathbf{q}\,, \tag{24}$$

where **q** is a vector whose elements are Gaussian noise with same standard deviation as the corresponding background errors.

(iv) Perform a 4D-Var inversion by minimizing

$$J(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{p}(y^{i} - H^{i}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(y^{i} - H^{i}\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{\mathrm{b}})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_{\mathrm{b}}), \tag{25}$$

where p is the number of observations.

4.1.1. *Finite-difference approximation*

In order to derive a full-rank estimate of the inverse Hessian for the CO_2 flux inversion, we reduced the dimension of the control vector by considering only the first 200 model grid cells with highest fossil fuel CO_2 emissions. Given this set-up, the dimensions of the control vector and the observational vector are $n = 200$ and $p = 33\,822$, respectively.

The following formula has then been used to estimate each element of the Hessian by finite-difference calculation:

$$\widehat{\mathbf{H}}_{i,j} = \frac{\{\nabla J(\mathbf{x} + \epsilon_i) - \nabla J(\mathbf{x})\}_j}{\epsilon}, \tag{26}$$

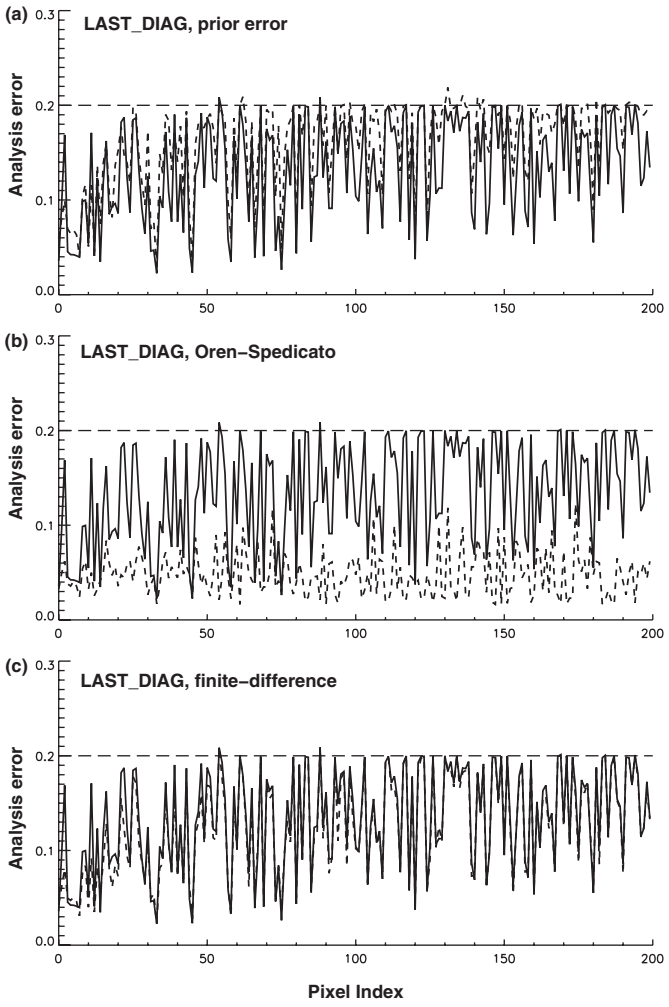

Figure 4. Analysis standard error estimates for all 200 optimized scaling factors using the finite-difference calculation (solid line) and the new LAST_DIAG BFGS diagonal preconditioner (dashed line): (a) prior error initialization, (b) Oren–Spedicato initialization, and (c) finite-difference estimate initialization. All BFGS calculations use the same 32 vector and gradient pairs obtained from one specific inversion.

Table 3. Performance statistics of the new LAST_DIAG BFGS preconditioning method with respect to the finite-difference inverse Hessian estimate.

BFGS LAST_DIAG (initialization)	No. of cycles	PCC	*a*, *b*	SDRE
Prior	1	0.85	0.73, 0.06	0.47
Oren–Spedicato	1	−0.06	−0.03, 0.05	0.61
Prior	60	0.81	0.68, 0.01	0.27
Oren–Spedicato	60	0.59	0.39, 0.03	0.35
Prior+ 3 inversions (BFGS_HYBRID)	60	0.94	0.91, 0.01	0.19

Column headings are as in Table 1.

where $\epsilon_i = (\epsilon\delta_{i,k})_k$, δ is the Dirac function and ϵ is a small real number. For this experiment $\epsilon = 0.01$, which corresponds to a 1% perturbation of the CO_2 flux for a particular grid cell. Since $\widehat{\mathbf{H}}$ is symmetric, this calculation requires $n + 1$ gradient calculations, which corresponds to $n + 1$ forward model integrations and $n + 1$ adjoint model integrations. Note that these gradient calculations can be performed in parallel. The resulting Hessian matrix must then be inverted in order to estimate the covariance matrix of analysis errors. Here we used the Linear Algebra PACKage (LAPACK) routines (http://www.netlib.org/lapack/; accessed 20 November 2014) to perform the inversion of $\widehat{\mathbf{H}}$ using LU decomposition.

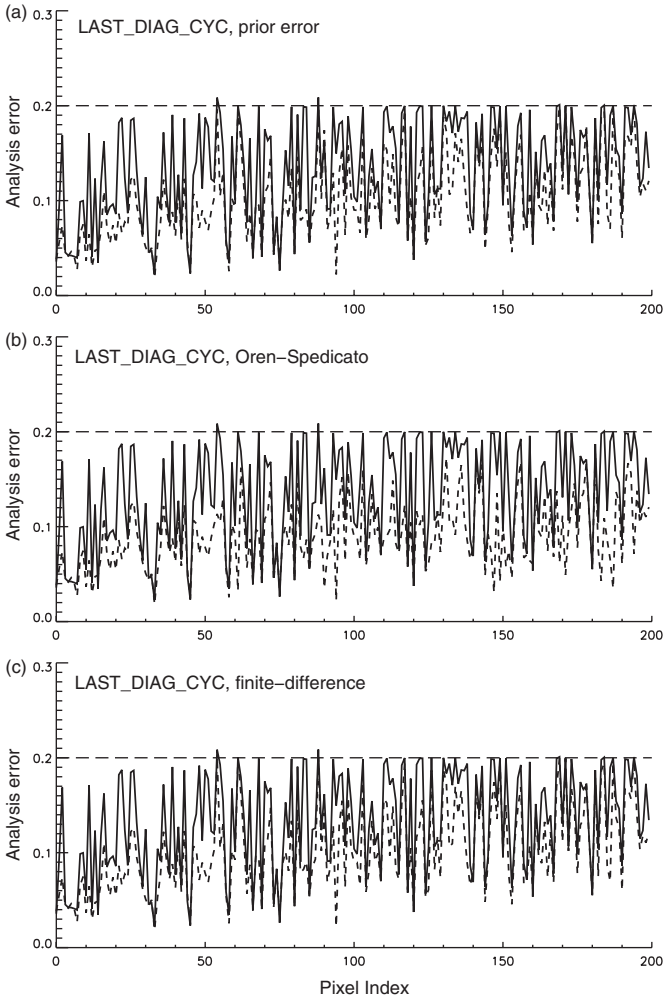

Figure 5. Analysis standard error estimates for all 200 optimized scaling factors using finite-difference calculation (solid line) and the new LAST_DIAG_CYC diagonal preconditioner using 60 cycles (dashed line): (a) prior error initialization, (b) Oren−Spedicato initialization, and (c) finite-difference initialization. All BFGS calculations use the same 32 vector and gradient pairs obtained from one specific inversion.

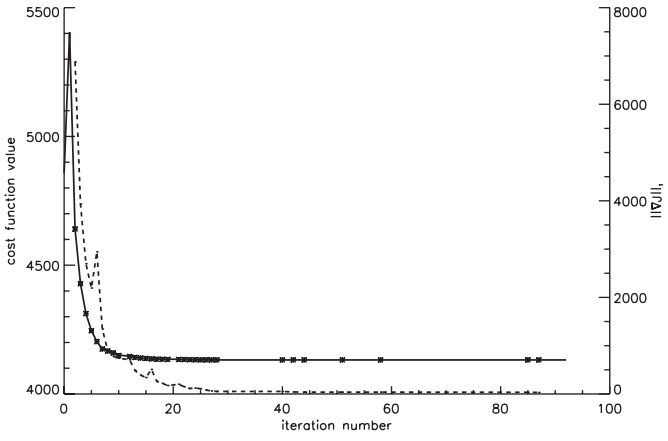

Figure 6. Cost function and norm of the gradient for 100 iterations of the BFGS minimization algorithm. Crosses represent accepted iterations.

4.1.2. *Stochastic estimates*

Figure 1 shows the analysis standard error for each scaling factor estimated using the finite-difference, Monte-Carlo, and gradient-based stochastic methods. Performance statistics for each method are summarized in Table 1. The Monte-Carlo estimate closely matches the finite-difference calculation, with a relative standard error of 10%, consistent with the theoretical value for an ensemble of 50 members (section 2.1). As explained in section 2.2, when using the gradient-based approach, there is no guarantee that 50 members will be sufficient to obtain a 10% relative standard error in the estimates. Here using 50 members results in a significant positive bias of about 50% in the estimate (Figure 1). Based on the Frobenius norm of the approximated inverse Hessian matrix, we found that 500 members were necessary to reach convergence. This yields an estimation of the analysis errors with similar performance as the Monte-Carlo estimate, with a bias of only 4%.

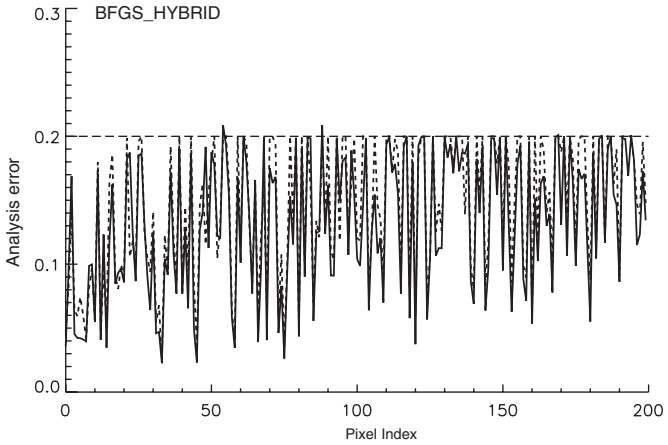

Figure 7. Analysis standard error estimates for all 200 optimized scaling factors using the finite-difference calculation (solid line) and the new hybrid approach (BFGS_HYBRID) (dashed line). As in previous figures, the calculation used 32 vector and gradient pairs obtained from one specific inversion.

It is well known that randomization methods produce spurious long-range error correlations due to sampling noise (Fisher and Courtier, 1995). Figure 2 shows the relative errors in the analysis-error correlation compared to the finite-difference estimate as a function of distance. The fact that the gradient-based approximation performs better at estimating the error correlations stems from the fact that only the observational term in the inverse Hessian formula (10) is stochastically approximated. In addition to significantly reducing the sampling noise, the gradient-based approximation requires much less computation than the Monte-Carlo estimate. In our experiment, about 100 BFGS line searches were necessary for each inversion in the Monte-Carlo ensemble, which corresponds to a total of 5000 gradient evaluations to estimate the analysis-error covariance matrix. The gradient-based estimate required only 500 gradient evaluations, which can all be performed in parallel.

4.1.3. *Existing BFGS preconditioners*

Figure 3 shows the analysis standard error for each scaling factor estimated using the finite-difference calculation and several existing BFGS algorithms described in section 3.2. Here INIT_B refers to the BFGS algorithm initialized with the diagonal matrix of prior errors instead of the Oren−Spedicato scalar (INIT_O). Performance statistics are reported in Table 2. Large errors and poor correlations are found for all approaches. The difference between the analysis errors calculated with the different methods largely reflects the influence of the initialization. Using INIT_B, only a few error reductions, albeit underestimated, are captured. They are generally associated with the highest error reductions calculated with finite difference. The INIT_O and LC_DIAG methods give similar results and show a significant underestimation of the analysis errors. As explained in section 3.2 (also Nocedal and Wright, 2006), in both methods the diagonal elements of the initial matrix tend to approximate the eigenspectra of the inverse Hessian. As a result, as seen on Figure 3, the analysis error estimates are good approximations to the lowest analysis errors, which are associated with the dominant eigenvalues of the Hessian.

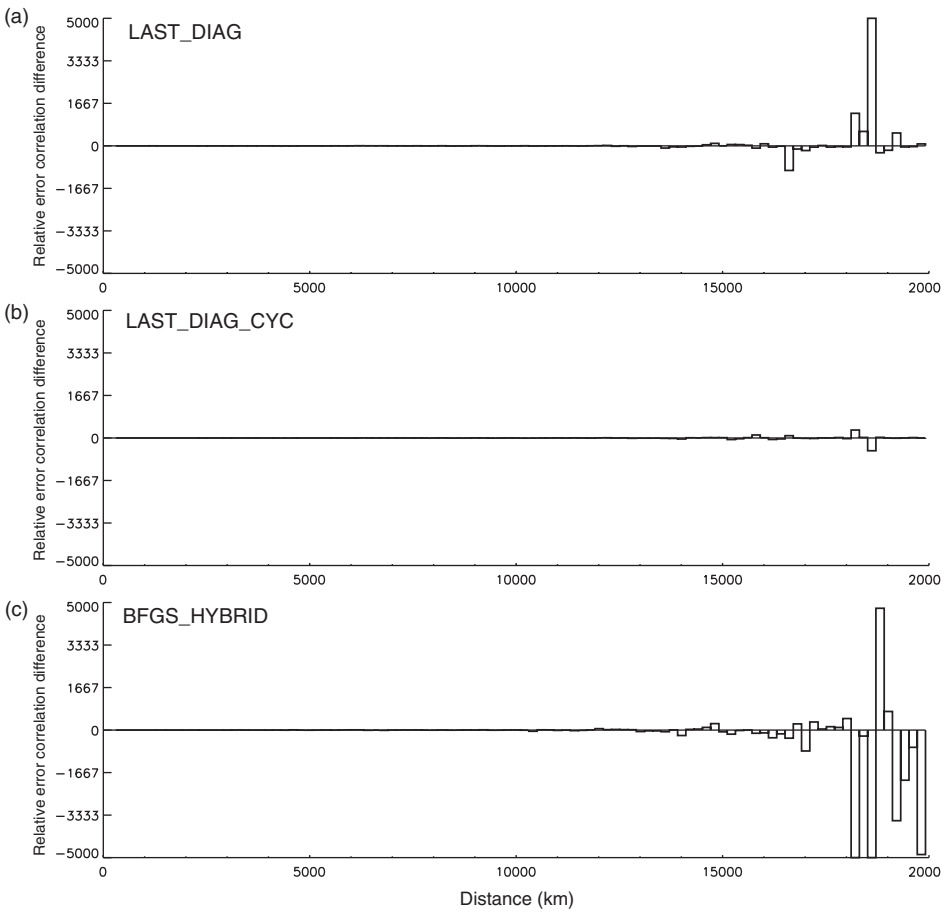

Figure 8. Relative differences between the analysis-error correlations calculated using finite-difference and the new BFGS algorithms as a function of distance (values have been averaged over all scaling factors for each distance bin): (a) LAST_DIAG, (b) LAST_DIAG_CYC, and (c) BFGS_HYBRID.

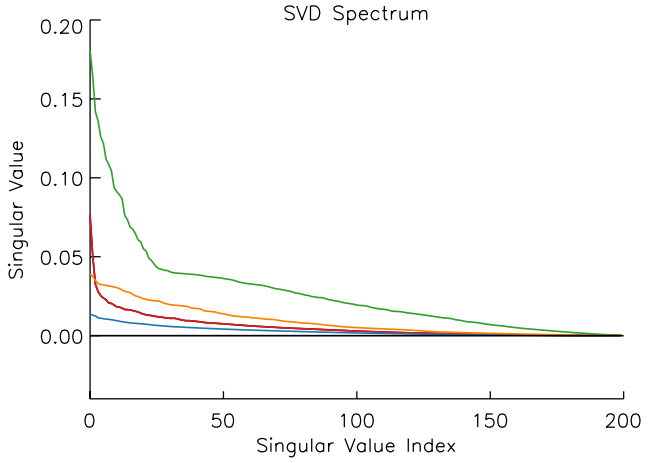

Figure 9. Singular values spectra of the inverse Hessian residual, which is the difference between the finite-difference estimate and the approximation ($\widehat{\mathbf{H}}_{\mathrm{FD}}^{-1} - \widehat{\mathbf{H}}_{\mathrm{approx}}^{-1}$) obtained with: Monte-Carlo (green), gradient-based randomization (blue), BFGS LAST_DIAG_CYC (orange), and BFGS_HYBRID (red) methods.

Our results also show that applying the GL_DIAG preconditioning method with prior error initialization yields a large number of analysis errors significantly higher than the prior error. Since by construction the inversion systematically reduces the prior error, these noisy values can be filtered out by simply resetting them to the prior error at each iteration. In this case we see that GL_DIAG give similar results as INIT_B, while significantly improving the magnitude of the error reduction.

Overall our results show that existing BFGS preconditioning techniques have little value for estimating the diagonal elements of the inverse Hessian matrix. An alternative approach is proposed in the next section.

4.1.4. *New BFGS preconditioner*

One weakness of the GL_DIAG preconditioner is that only element-wise products of $\mathbf{y}_k$ and $\mathbf{s}_k$ are considered when updating the diagonal matrix, neglecting the influence of cross-product terms. In addition, the diagonal preconditioner is updated independently from the main inverse Hessian update. Here we propose to improve upon this method by using the LAST_DIAG preconditioner defined by Eq. (21), which ensures that the information brought by each inverse Hessian update is also used to update the preconditioner. This feedback mechanism dramatically improves the accumulation of information along the diagonal, as shown in Figure 4. Table 3 summarizes the performance of this preconditioning technique. Results are improved compared to previous BFGS algorithms, whether the prior or the Oren-Spedicato scalar are used for initialization (section 4.1.3). Also shown are the analysis errors obtained by initializing LAST_DIAG with the finite-difference estimate. Although the relative stability of the 'true' initial diagonal matrix under BFGS updates may be interpreted as a desirable property of the algorithm, in fact it essentially reflects the strong influence of the initialization, as we will show in the next section.

From these results it is evident that although the new preconditioning technique significantly improves the performance of the BFGS algorithm compared to previous approaches, the influence of the initialization on the estimate is still dominant. In the next section we discuss this issue and propose a new method to address this problem.

4.1.5. *Cycling the BFGS*

As shown previously, inverse Hessian approximations calculated using the BFGS algorithm are strongly dependent upon the initialization. Among all approaches, the LAST_DIAG algorithm

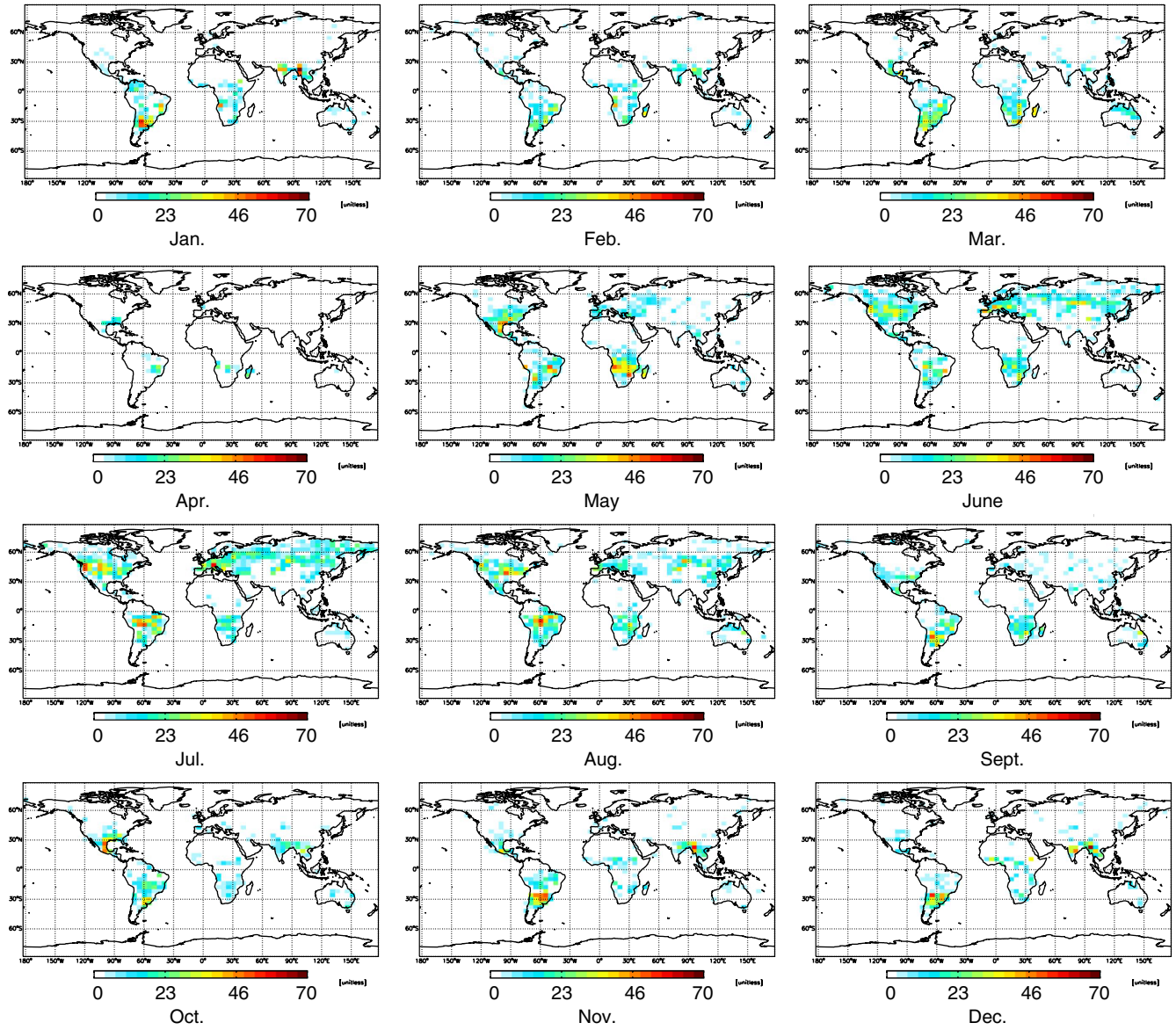

Figure 10. Monte-Carlo estimate (50 members) of the relative error reduction (% of prior error) in biospheric fluxes for each month of the year 2010.

is the least sensitive to the initialization. The reason is that each update of the initial diagonal matrix ($\widehat{\mathbf{H}}_k^0$) reuses information from all previous vector and gradient pairs multiple times. More precisely, at iteration k, the LAST_DIAG algorithm has used the pair $(\mathbf{y}_m, \mathbf{s}_m)$ exactly $k - m + 1$ times to update the inverse Hessian. Therefore the amount of information coming from the gradient and vector pairs is greatly enhanced and acts as a forgetting factor with respect to the initial matrix ($\widehat{\mathbf{H}}_0^0$).

A direct extension of this approach is to continue iterating the algorithm after all available pairs have been used and perform several cycles. Note that, although each cycle reuses information from the same pairs, it is initialized using a different inverse Hessian diagonal. Figure 5 shows the analysis errors calculated using 60 cycles of LAST_DIAG for different initializations (prior error, Oren–Spedicato, and finite-difference estimate). All methods now give similar results and reduce the relative standard error in the analysis error compared to the case with only one cycle. This demonstrates the need for cycling the BFGS algorithm in order to obtain both convergence (independence from the initialization) and better estimates of the diagonal elements of the inverse Hessian. The number of cycles chosen here (60) was based on the convergence of the Frobenius norm of the BFGS inverse Hessian approximation (not shown). Note that this number may vary with the dimension of the problem, since we found that convergence can be achieved after only 30 cycles for a similar experiment with a control vector of dimension 70. In the following sections we will refer to the LAST_DIAG algorithm with cycling as LAST_DIAG_CYC.

4.1.6. *Hybrid approach*

Although the LAST_DIAG_CYC method consisting of cycling the LAST_DIAG algorithm proved to be superior over all other BFGS methods for approximating the analysis errors, it still gives poorer performances than the stochastic approaches. This is due to the fact that in practice the dimension of the optimization subspace from which the inverse Hessian approximation is built is much smaller than the dimension of the entire control space (n). As a result, the inverse Hessian estimate has a rank much lower than n. In the case when line searches are not perfect (this study), the rank of the inverse Hessian estimate is further degraded. Even when n is relatively small, it may not be practical to perform enough iterations to span the entire control space. Figure 6 shows the cost function value and the norm of the gradient for the first 100 line searches of the BFGS minimization algorithm. After 25 accepted iterations, the proportion of accepted iterations drops dramatically (only 7 out of 68). One approach to increase the rank of the BFGS inverse Hessian approximation is to use gradient and vector pairs from several inversions taken from a Monte-Carlo ensemble. Since the perturbations applied

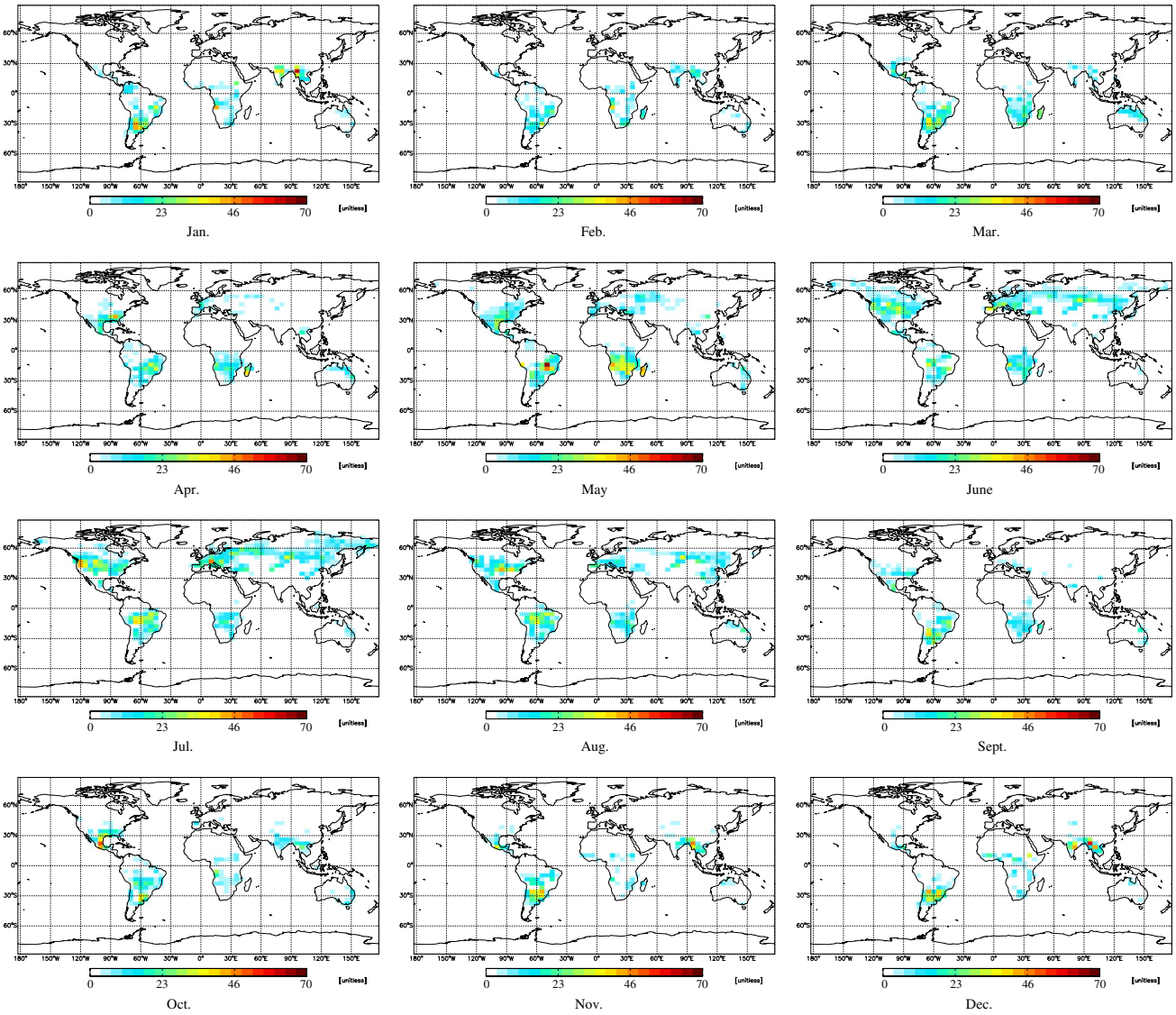

Figure 11. Gradient-based estimate (1800 members) of the relative error reduction (% of prior error) in CO_2 biospheric fluxes for each month of the year 2010.

to the prior and the observations are different between inversions (section 2.1), in practice the search directions generated during the minimization will be different and allow for sampling a larger subspace of the control space. While directions corresponding to strongest error reductions tend to be similar among inversions, directions associated with smaller error reductions show more variability. Figure 7 shows the results obtained with this hybrid approach (BFGS_HYBRID), which consists of applying the LAST_DIAG_CYC algorithm to the set of gradient and vector pairs from three different Monte-Carlo ensemble members. The filtering described in section 4.1.3 was also applied throughout the process to correct for potential prior error increases. Performance statistics are summarized Table 3. The BFGS_HYBRID method dramatically improves the analysis error estimates, with results now similar to the stochastic methods.

Figure 8 shows the relative error in the analysis-error correlations with respect to the finite-difference estimate using the LAST_DIAG, LAST_DIAG_CYC, and BFGS_HYBRID methods. Cycling the BFGS (LAST_DIAG_CYC) remove the spurious long-range correlations initially produced (LAST_DIAG). However, combining gradient and vector pairs from different inversions in the BFGS estimate (BFGS_HYBRID) introduces significant spurious long-distance correlations. Overall, these results demonstrate that the BFGS_HYBRID approach provides analysis-error estimates similar to the stochastic methods, while mitigating the error correlations sampling noise obtained with the Monte-Carlo estimate.

Note that the idea of using information from multiple inversions to improve the inverse Hessian estimate is not new. For example, the LMP described in Tshimanga *et al.* (2008) and Gratton *et al.* (2011a), when used in the context of incremental 4D-Var data assimilation, can construct a BFGS approximation of the inverse Hessian from the sequence of quadratic CG minimizations of the outer loops, gradually improving the preconditioning of the systems. In the BFGS_HYBRID approach, the ensemble of minimization problems is obtained through random perturbations of the prior and observations in the 4D-Var cost function. An important underlying assumption is that the forward model is approximately linear so that the Hessian of the cost function does not depend on the particular prior or observations considered. However, different starting points for the prior and the observations in the minimization will result in different subspaces of the control space being explored. It is also worthwhile to note that the EIH method presented in Gejadze *et al.* (2011) also combines information from several BFGS inverse Hessian estimates using stochastic methods. Nevertheless, in their approach the ensemble of inverse Hessian estimates is obtained from different quadratic problems (unlike ours), which aims at taking into account large nonlinearities in their inverse problem.

4.1.7. Singular value decomposition analysis

A singular vector decomposition (SVD) analysis provides a useful framework to assess the overall performance of the different algorithms employed to approximate the inverse Hessian matrix

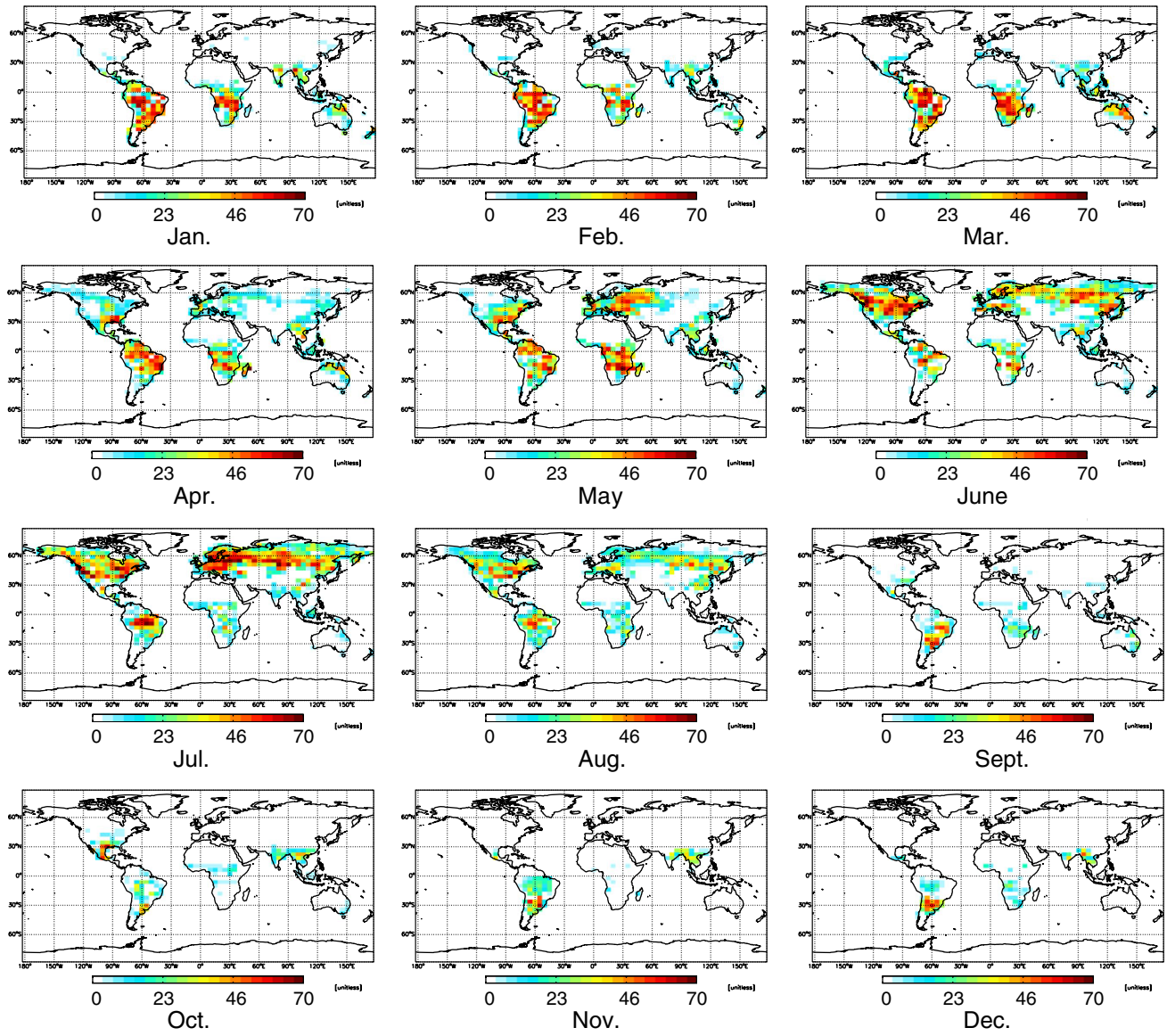

Figure 12. BFGS LAST_DIAG_CYC estimate of the relative error reduction (% of prior error) in CO_2 biospheric fluxes for each month of the year 2010.

(see Veersé *et al.*, 2000). Figure 9 shows the singular value spectra of the inverse Hessian 'residual' (the difference between the finite-difference calculation and the approximate inverse Hessian) for the stochastic approaches and the new BFGS-based algorithms, namely BFGS LAST_DIAG_CYC and BFGS_HYBRID. This demonstrates the superiority of the BFGS_HYBRID and gradient-based approaches over other techniques, and clearly shows the degrading effect of the sampling noise in the Monte-Carlo estimate, for which significant biases are found in the higher part of the spectra. This is expected since spurious long-range error correlations will translate into increased variances error for the eigenvectors of the inverse Hessian.

4.2. *Application to a high-dimensional problem*

In this OSSE experiment we use the Carbon Monitoring System (CMS) flux estimation model to infer error reduction in global monthly CO_2 biospheric fluxes from the assimilation of ACOS-GOSAT X_{CO_2} data in 2010 (Liu *et al.*, 2014). More specifically, the control vector is composed of all CO_2 biospheric flux scaling factors at every grid cell of a $4^\circ \times 5^\circ$ global simulation, for every month. It is therefore of dimension $72 \times 46 \times 12 = 39\,744$. With such a large number of scaling factors, it is not possible to perform a finite-difference estimate of the analysis-error variance. However, since the relative standard error in the Monte-Carlo estimate is $\leq 10\%$ with 50 ensemble members, independently from the dimension of the problem, we will consider this approximation as a reference against which other methods will be evaluated. In this study, the prior error covariance matrix **B** is defined as diagonal (therefore assuming no error correlation between fluxes), and was constructed from a Monte Carlo run of CASA-GFED 3 by sampling the distributions of model parameters. For more details on the experimental set-up, the reader is invited to refer to Liu *et al.* (2014).

Figures 10–13 represent maps of monthly error reduction relative to the prior error after inversion of the CO_2 biospheric fluxes, calculated using the Monte-Carlo, gradient-based, BFGS LAST_DIAG_CYC, and BFGS_HYBRID approaches. Performance statistics from these results are summarized in Figure 14, which shows a scatter plot of each approximation against the Monte-Carlo estimate, as well as the associated linear regression analysis. Generally the Monte-Carlo and gradient-based estimates show very similar error reduction patterns and magnitudes. However, the Monte-Carlo method tends to exhibit higher error reduction than the gradient-based approach over some areas with weak constraints on the fluxes (i.e. Europe and Asia in September). The very good agreement between the two methods is characterized by a strong correlation coefficient ($R = 0.93$) and a linear regression fit close to 1 : 1. Note that the size of the random gradient ensemble necessary to obtain convergence of the error reduction estimates is significantly greater for this realistic inversion (1800 members) than for the previous low-dimensional problem (500

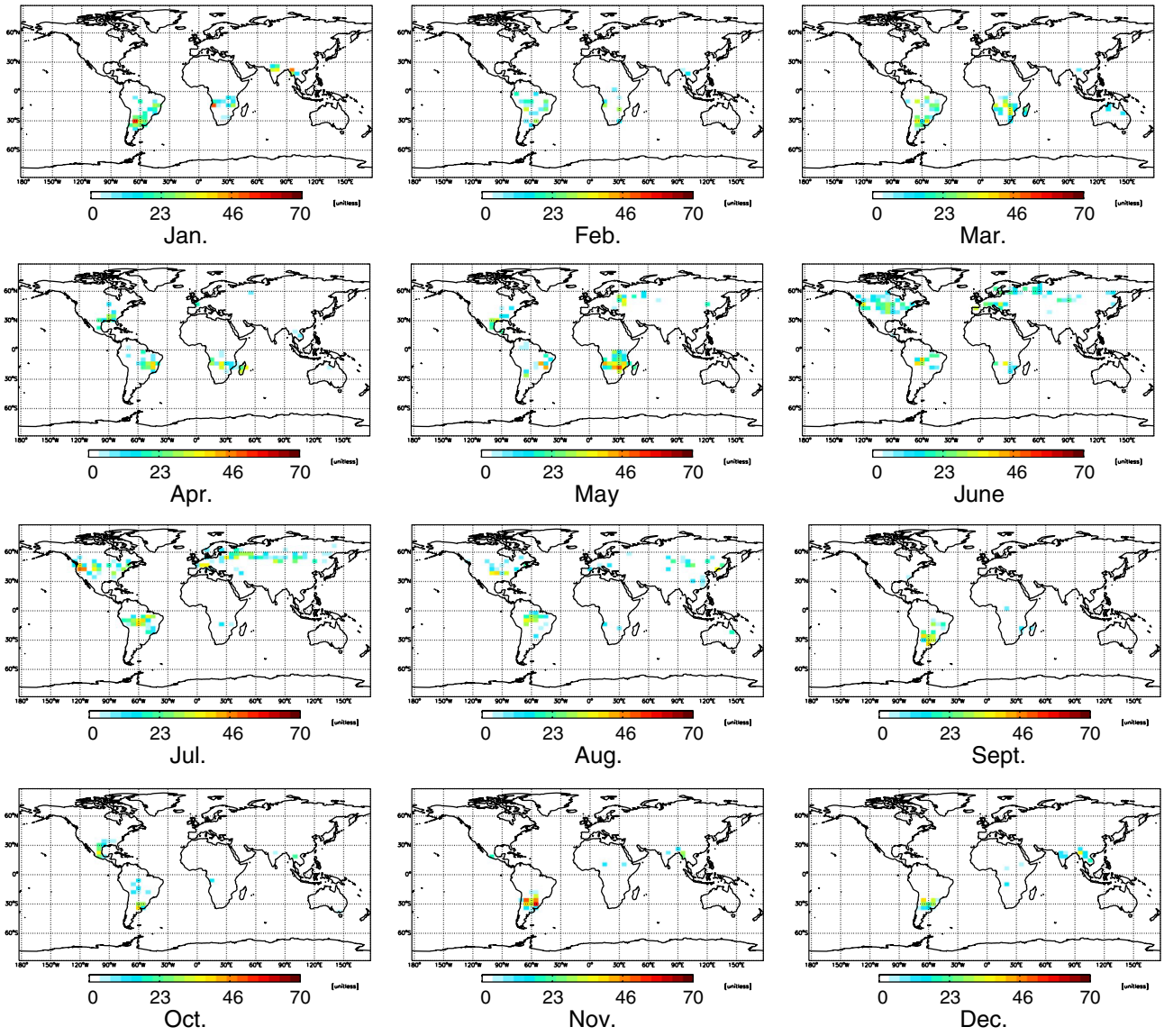

Figure 13. BFGS_HYBRID estimate (five members) of the relative error reduction (% of prior error) in CO_2 biospheric fluxes for each month of the year 2010.

members). Although the performances of the stochastic estimates are by nature independent from the dimension of the problem, the convergence rate of the gradient-based method (as opposed to the Monte-Carlo one) can still be affected by other characteristics of the problem (see section 3.3.2). In particular, the realistic inversion is more ill-posed, which results in lower error reductions and higher analysis error correlations than in the low-dimensional experiment, and likely explains the observed differences in performance for the gradient-based estimate.

Similar to results obtained with the previous low-dimensional experiment, the analysis errors estimated with the BFGS LAST_DIAG_CYC algorithm using information from only one inversion yields a significant overestimation of the error reduction compared to the Monte-Carlo estimate. Although the spatial distribution of the error reduction is well reproduced in a relative sense (compare Figures 10 and 12), which results in a good correlation between the two methods ($R = 0.60$), the linear regression analysis (Figure 14) shows a consistent overestimation of the error reduction by up to a factor 3 by the BFGS method. Again, using information from multiple Monte-Carlo inversions to compute the BFGS estimate (BFGS_HYBRID) results in a dramatic improvement of the analysis error approximation, as seen from Figure 13 and the linear regression fit (Figure 14). In particular, the average BFGS overestimation of the error reduction (50%) is much smaller than for the BFGS LAST_DIAG_CYC estimate. Overall, these results show that, while the performance of the gradient-based method is fairly unaffected by the dimension of the problem, the accuracy of the improved BFGS estimates can be significantly reduced for high-dimensional inversions. Still, the BFGS_HYBRID approach offers good scalability properties. Here five Monte-Carlo members were used to compute the inverse Hessian, while three members were used for the low-dimensional case. Therefore, using a similar number of Monte-Carlo samples in a system whose dimension is two orders of magnitude higher than the low-dimensional case results in a decrease of only 0.20 in the linear regression coefficient. Although using more Monte-Carlo ensemble members in the BFGS_HYBRID algorithm could in principle improve the performance of the estimate, the computational cost of the calculation can become rapidly prohibitive when more than 200 gradient and vector pairs are considered (see A.1.2). This is primarily due to the numerous BFGS cycles through the gradient and vector pairs necessary to efficiently accumulate information on the diagonal of the inverse Hessian. As an example, for our BFGS_HYBRID estimate of the inverse Hessian using five Monte-Carlo inversions (i.e. 140 gradient and vector pairs), performing one BFGS cycle takes about 5.5 h on a system with dual hexa-core 2.67 GHz Intel Xeon processors and 12 MB of RAM.

5. Conclusions

In this study we tested several numerical methods to estimate the analysis-error covariance matrix of high-dimensional 4D-Var inverse problems, when the forward model considered is

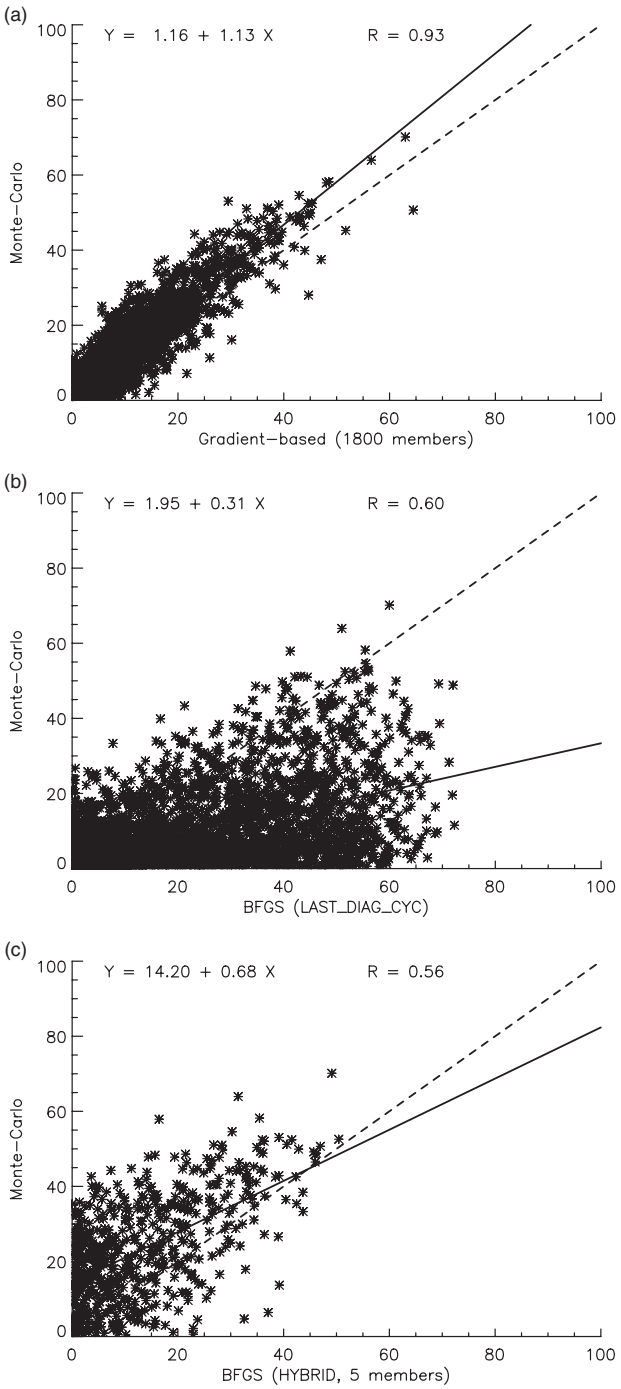

Figure 14. Scatter plot and linear regression analysis for the comparison between the Monte-Carlo and (a) the gradient-based, (b) BFGS LAST_DIAG_CYC and (c) BFGS_HYBRID methods. The dashed line denotes 1:1.

approximately linear. An experimental framework consisting of a low-dimensional inversion of CO_2 emissions constrained by remote-sensing pseudo-observations of CO_2 columns has been used to derive a direct finite-difference estimate of the inverse Hessian of the cost function. Both stochastic and deterministic approaches were then tested and evaluated against the finite-difference estimate.

A standard Monte-Carlo method as well as a calculation based on stochastic gradients of the cost function gave excellent analysis-error estimates, with a relative standard uncertainty within 10% of the finite-difference estimate. Spurious long-range error correlations were produced with both approaches due to sampling noise. While mitigating the sampling noise, the gradient-based method required about ten times fewer gradient evaluations than the Monte-Carlo calculation.

Deterministic estimations of the inverse Hessian have also been tested using the BFGS algorithm. Existing preconditioning methods were unable to provide useful estimates of the analysis errors (relative error $> 120\%$). Therefore, a new diagonal preconditioner has been proposed (LAST_DIAG), which better accumulates information on the diagonal of the inverse Hessian and dramatically improves the results (relative error $< 50\%$). We found that the BFGS method was very sensitive to initialization and that cycling the algorithm a number of times (~ 40) using the same gradient and vector pairs (LAST_DIAG_CYC) was necessary to reach convergence and obtain more accurate analysis-error estimates (relative error $< 30\%$). However, Hessian information obtained from only one inversion yielded analysis-error estimates of poorer quality than obtained with the stochastic methods. We showed that using a hybrid approach (BFGS_HYBRID) which combines BFGS algorithm cycles with inversion outputs from only a few (3 here) Monte-Carlo ensemble members, it is possible to obtain analysis-error estimates of similar quality to the stochastic calculations.

These methods have also been tested with a realistic high-dimensional problem, consisting of a global inversion of monthly CO_2 biospheric fluxes for the year 2010, using the Carbon Monitoring System (CMS) Flux Pilot Project developed at NASA (Liu *et al.*, 2014). In this case the gradient-based randomization approach and the improved BFGS algorithms were compared to a Monte-Carlo estimate with 50 members, and only error variances were considered. The gradient-based estimate required about three times more sampling to reach convergence than in the low-dimensional experiment. These results demonstrate that unlike for the Monte-Carlo approach, the computational efficiency of the gradient-based method can significantly vary between inverse problems, and suggest more sampling will be necessary as the average error reduction of the inversion decreases. The improved BFGS_HYBRID algorithm exhibited good scalability with respect to the dimension of the problem. Indeed, using a similar number of ensemble members (5) as for the low-dimensional case (3), good correlations ($R = 0.56$) with respect to the standard Monte-Carlo calculation were obtained, as well as a good spatial representation of the main error variance characteristics.

In summary, the choice of the algorithm employed should be guided by the targeted application and by the computational resources available. A SVD analysis revealed that overall best performances are obtained using the gradient-based randomization approach. Although in our experiments 500 to 1800 gradient evaluations were necessary to converge, in practice these calculations can be all performed in parallel. Therefore, this method is especially well suited when an evaluation of the analysis-error covariances is sought prior to optimization. However, it is worthwhile to note that the gradient-based approach can be superseded by the Monte-Carlo method in cases where the optimization requires only a small number of iterations. On the other hand, the improved BFGS cycling approach (LAST_DIAG_CYC) affords a relatively good approximation of the inverse Hessian in directions corresponding to highest error reductions, at a small computational cost. While it may not be adapted to fine-scale error analysis, other applications requiring fast evaluations of the inverse Hessian may be envisaged. For instance, it would be useful to test this new preconditioner in a limited-memory BFGS minimization context in order to assess its impact on the convergence rate. Moreover, in an incremental 4D-Var assimilation system, an approximate Hessian derived from a similar methodology could be used to precondition the CG minimization of the inner loops. In particular, the LAST_DIAG_CYC preconditioner could be used in the quasi-Newton LMP approach proposed by Tshimanga *et al.* (2008) and Gratton *et al.* (2011a) to further improve the convergence of the inner loops. Another interesting application of the BFGS method in variational data assimilation was suggested by Fisher and Courtier (1995). In the case where observations occur only at or after the time at which the forecast error matrix is required, the BFGS transformation formula can be used to propagate the analysis-error covariance matrix using only information

generated during the minimization, allowing a computationally efficient approximate Kalman filter to be implemented. Such implementation should benefit from the BFGS LAST_DIAG_CYC algorithm proposed, by improving the inverse Hessian estimate. Finally, the hybrid approach (BFGS_HYBRID) combines an accurate estimate of the analysis-error covariance matrix with the numerical efficiency of the BFGS matrix–vector product recursive algorithm (see Appendix). Therefore, its use would be particularly well adapted to characterize the information content of high-dimensional inversions using SVD algorithms, since they usually require numerous matrix–vector products. In addition, since the performance of the BFGS HYBRID estimate seems only weakly impacted by the dimension and information characteristics of the problem, this algorithm can be an efficient alternative to either the Monte-Carlo or the gradient-based analysis error estimates in case numerous iterations are required in the optimization and a large number of random gradients is necessary for the inverse Hessian estimate to converge.

Acknowledgements

This work has been funded by the NASA CMS program (grant number NNX12AR31G), the NASA Atmospheric CO_2 Observations from Space program (grant number NNX10AT42G), as well as the NASA New Investigator Program NNX10AR06G. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. The first author would also like to thank Richard Byrd for a number of useful discussions about the BFGS minimization method.

Appendix

Practical algorithms for high-dimensional systems

A1. BFGS method

A1.1. Matrix-vector product

The implicit formulation in Eq. (15) can be used to derive a two-loop recursion algorithm to calculate the product $\widehat{\mathbf{H}}_k\mathbf{v}$ of the approximate inverse Hessian with any vector $\mathbf{v}$ (Nocedal, 1980):

Algorithm 1 (L-BFGS two-loop recursion)

$\mathbf{q} \leftarrow \mathbf{v}$
for $i = k-1$ to $k-2, \ldots, k-m$ **do**
  $\alpha_i \leftarrow \rho_i \mathbf{s}_i^{\mathrm{T}}\mathbf{q}$
  $\mathbf{q} \leftarrow \mathbf{q} - \alpha_i \mathbf{y}_i$
end for
$\mathbf{r} \leftarrow \widehat{\mathbf{H}}_k^0\mathbf{q}$
for $i = k-m$ to $k-m+1, \ldots, k-1$ **do**
  $\beta \leftarrow \rho_i \mathbf{y}_i^{\mathrm{T}}\mathbf{r}$
  $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{s}_i(\alpha_i - \beta)$
end for
stop with result $\widehat{\mathbf{H}}_k\mathbf{v} = \mathbf{r}$

where m corresponds to the number of gradient and vector pairs stored in computer memory. Therefore, the two-loop recursion algorithm requires $4mn + n$ multiplications if $\widehat{\mathbf{H}}_k^0$ is diagonal, where n represents the dimension of the control state.

A1.2. Matrix element extraction

In order to be able to calculate specific elements of $\widehat{\mathbf{H}}_k$, the formula (12) can be rewritten in the form (Fisher and Courtier, 1995):

$$\widehat{\mathbf{H}}_{k+1} = \widehat{\mathbf{H}}_k + \mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}\left(\frac{1}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k} + \frac{\mathbf{y}_k^{\mathrm{T}}\widehat{\mathbf{H}}_k\mathbf{y}_k}{(\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k)^2}\right) - \frac{1}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k}\left\{\mathbf{s}_k\left(\widehat{\mathbf{H}}_k^{\mathrm{T}}\mathbf{y}_k\right)^{\mathrm{T}} + \left(\widehat{\mathbf{H}}_k\mathbf{y}_k\right)\mathbf{s}_k^{\mathrm{T}}\right\}, \tag{A1}$$

where for any vectors $\mathbf{u}$ and $\mathbf{v}$, $(uv^{\mathrm{T}})_{i,j} = u_i v_j$. Since $\widehat{\mathbf{H}}_k$ is symmetric, one has $\widehat{\mathbf{H}}_k^{\mathrm{T}} = \widehat{\mathbf{H}}_k$, and the product $\widehat{\mathbf{H}}_k\mathbf{y}_k$ in Eq. (A1) can be computed using Algorithm 1. When m gradient and vector pairs are used and the diagonal initial matrix $(\widehat{\mathbf{H}}_k^0)$ is updated at each iteration k, the algorithm requires

$$\sum_{q=1}^{m}\sum_{i=1}^{q}(4in + 2n) = n(2m^3 + 9m^2 + 11m)$$

multiplications.

A2. Gradient-based randomization

A2.1. Matrix-vector product

As explained in section 2.2, the formula (10) needs to be inverted in order to compute the analysis error covariance matrix. This can be done iteratively using the Sherman–Morrison–Woodbury formula. For any invertible matrix $\mathbf{M}$, and any vectors $\mathbf{u}$, $\mathbf{v}$, one has:

$$(\mathbf{M} + \mathbf{u}\mathbf{v}^{\mathrm{T}})^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{u}}. \tag{A2}$$

Applying Eq. (A2) to Eq. (10) gives the following recursive algorithm to compute the analysis-error covariance matrix:

$$\left.\begin{aligned}\mathbf{P}_{k+1}^{\mathrm{a}} &= \mathbf{P}_k^{\mathrm{a}} - \frac{(1/m)\mathbf{P}_k^{\mathrm{a}}\nabla J(\mathbf{x}_{\mathrm{b}})^k\{\nabla J(\mathbf{x}_{\mathrm{b}})^k\}^{\mathrm{T}}\mathbf{P}_k^{\mathrm{a}}}{1 + (1/m)\{\nabla J(\mathbf{x}_{\mathrm{b}})^k\}^{\mathrm{T}}\mathbf{P}_k^{\mathrm{a}}\nabla J(\mathbf{x}_{\mathrm{b}})^k},\\ \mathbf{P}_0^{\mathrm{a}} &= \mathbf{B},\end{aligned}\right\} \tag{A3}$$

where $\nabla J(\mathbf{x}_{\mathrm{b}})^k$ denotes the kth sample of the random variable $\nabla J(\mathbf{x}_{\mathrm{b}})$ (section 2.2), and m is the number of samples used.

From Eq. (A3) we obtain the following recursive formula to compute the product $\mathbf{P}^{\mathrm{a}}\mathbf{q}$ of the analysis-error covariance matrix with any vector $\mathbf{q}$:

$$\left.\begin{aligned}\mathbf{u}_{k+1} &= \mathbf{u}_k - \frac{\mathbf{v}_k^k(\mathbf{v}_k^k)^{\mathrm{T}}\mathbf{q}}{1 + (1/\sqrt{m})\{\nabla J(\mathbf{x}_{\mathrm{b}})^k\}^{\mathrm{T}}\mathbf{v}_k^k},\\ \mathbf{v}_{k+1}^{k+i} &= \mathbf{v}_k^{k+i} - \frac{1}{\sqrt{m}}\frac{\mathbf{v}_k^k(\mathbf{v}_k^k)^{\mathrm{T}}\nabla J(\mathbf{x}_{\mathrm{b}})^{k+i}}{1 + (1/\sqrt{m})\{\nabla J(\mathbf{x}_{\mathrm{b}})^k\}^{\mathrm{T}}\mathbf{v}_k^k}, \quad 1 \le i \le m-k,\\ \mathbf{u}_0 &= \mathbf{B}\mathbf{q},\\ \mathbf{v}_0^i &= \frac{1}{\sqrt{m}}\mathbf{B}\,\nabla J(\mathbf{x}_{\mathrm{b}})^i, \qquad 0 \le i \le m,\end{aligned}\right\} \tag{A4}$$

where m is the total number of gradient samples used, and $\mathbf{u}_m$ gives the approximated product $\mathbf{P}^{\mathrm{a}}\mathbf{q}$.

At each iteration $\mathbf{u}_k$ and $\mathbf{v}_k^{k+i}(0 \le i \le m-k)$ are calculated and stored for the next iteration. The initial terms $\mathbf{u}_0$, $\mathbf{v}_0^i(0 \le i \le m)$ are calculated using the algorithms presented in Singh *et al.* (2011), based on a computationally efficient square-root formulation for $\mathbf{B}$. If $\mathbf{B}$ is diagonal, the total number of multiplications required to approximate the product $\mathbf{P}^{\mathrm{a}}\mathbf{q}$ using m gradient samples is $nm^2 + 2nm - n$.

A2.2. Matrix element extraction

Once the vectors $\mathbf{v}_k^k$ have been computed using Eq. (A4), any element of $\mathbf{P}^{\mathrm{a}}$ can be approximated using the following recursive formula:

$$\left.\begin{aligned}(\mathbf{P}^{\mathrm{a}}_{k+1})_{i,j} &= (\mathbf{P}^{\mathrm{a}}_{k})_{i,j} - \frac{(\mathbf{v}_k^k)_i(\mathbf{v}_k^k)_j}{1+(1/\sqrt{m})\{\nabla J(\mathbf{x}_{\mathrm{b}})^k\}^{\mathrm{T}}\mathbf{v}_k^k},\\ (\mathbf{P}^{\mathrm{a}}_{0})_{i,j} &= \mathbf{B}_{i,j}\,.\end{aligned}\right\} \tag{A5}$$

If $\mathbf{B}$ is diagonal, the total number of multiplications required to approximate any elements $(\mathbf{P}^{\mathrm{a}})_{i,j}$ using m gradient samples is nm^2.

References

Anderson J. 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**: 2884–2903.

Bannister RN. 2008. A review of forecast-error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast-error covariances. *Q. J. R. Meteorol. Soc.* **134**: 1951–1970.

Berre L, Desroziers G. 2010. Filtering of background-error variances and correlations by local spatial averaging: A review. *Mon. Weather Rev.* **138**: 3693–3720.

Bonnans JF, Gilbert JC, Lemaréchal C, Sagastizábal CA. 2006. *Numerical Optimization: Theoretical and Practical Aspects*. Springer: Berlin.

Broyden C. 1969. A new double-rank minimisation algorithm. Preliminary report. *Am. Math. Soc. Notices* **16**: 670.

Buehner M. 2012. Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Mon. Weather Rev.* **140**: 617–636.

Chevallier F, Breon FM, Rayner PJ. 2007. Contribution of the orbiting carbon observatory to the estimation of CO_2 sources and sinks: Theoretical study in a variational data assimilation framework. *J. Geophys. Res. (Atmos.)* **112**: D09307, doi: 10.1029/2006JD007375.

Chevallier F, Wang T, Ciais P, Maignan F, Bocquet M, Arain MA, Cescatti A, Chen J, Dolman AJ, Law BE, Margolis HA, Montagnani L, Moors EJ. 2012. What eddy-covariance measurements tell us about prior land flux errors in CO_2 flux inversion schemes. *Global Biogeochem. Cycles* **26**: GB1021, doi: 10.1029/2010GB00397.

Courtier P, Thepaut J, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387.

Daescu DN. 2008. On the sensitivity equations of four-dimensional variational (4D-Var) data assimilation. *Mon. Weather Rev.* **136**: 3050–3065.

Davidon W. 1991. Variable metric method for minimization. *SIAM J. Optim.* **1**: 1–17.

Deng F, Jones DBA, Henze DK, Bousserez N, Bowman KW, Fisher JB, Nassar R, O'Dell C, Wunch D, Wennberg PO, Kort EA, Wofsy SC, Blumenstock T, Deutscher NM, Griffith DWT, Hase F, Heikkinen P, Sherlock V, Strong K, Sussmann R, Warneke T. 2014. Inferring regional sources and sinks of atmospheric CO_2 from GOSAT XCO_2 data. *Atmos. Chem. Phys.* **14**: 3703–3727.

Dennis J, More J. 1977. Quasi-Newton methods, motivation and theory. *SIAM Rev.* **19**: 46–89.

Dennis J, Wolkowicz H. 1993. Sizing and least-change secant methods. *SIAM J. Numer. Anal.* **30**: 1291–1314.

Desroziers G, Brousseau P, Chapnik B. 2005. Use of randomization to diagnose the impact of observations on analyses and forecasts. *Q. J. R. Meteorol. Soc.* **131**: 2821–2837.

Enting I. 2002. *Inverse Problems in Atmospheric Constituent Transport*, Vol. 1648. Cambridge University Press: Cambridge, UK.

Evensen G. 2007. *Data Assimilation*. Springer: Berlin.

Fisher M, Courtier P. 1995. *Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data Assimilation, Technical Memorandum* 45. ECMWF: Reading, UK.

Gejadze IY, Copeland GJM, Le Dimet F, Shutyaev V. 2011. Computation of the analysis-error covariance in variational data assimilation problems with nonlinear dynamics. *J. Comput. Phys.* **230**: 7923–7943.

Gilbert J, Lemaréchal C. 1989. Some numerical experiments with variable-storage quasi-Newton algorithms. *Math. Program.* **45**: 407–435.

Gratton S, Laloyaux P, Sartenaer A, Tshimanga J. 2011a. A reduced and limited-memory preconditioned approach for the 4D-Var data assimilation problem. *Q. J. R. Meteorol. Soc.* **137**: 452–466.

Gratton S, Sartenaer A, Tshimanga J. 2011b. On a class of limited memory preconditioners for large-scale linear systems with multiple right-hand sides. *SIAM J. Optim.* **21**: 912–935.

Henze DK, Hakami A, Seinfeld JH. 2007. Development of the adjoint of GEOS-Chem. *Atmos. Chem. Phys.* **7**: 2413–2433.

Houtekamer P, Mitchell H. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137.

Le Dimet F, Navon I, Daescu D. 2002. Second-order information in data assimilation. *Mon. Weather Rev.* **130**: 629–648.

Le Dimet F, Shutyaev V. 2005. On deterministic error analysis in variational data assimilation. *Nonlin. Proc. Geophys.* **12**: 481–490.

Leong WJ, Chen C. 2013. A class of diagonal preconditioners for limited memory BFGS method. *Optim. Methods Softw.* **28**: 379–392.

Liu J, Bowman KW, Lee M, Henze DK, Bousserez N, Brix H, Collatz GJ, Menemenlis D, Ott L, Pawson S, Jones D, Nassar R. 2014. Carbon monitoring system flux estimation and attribution: Impact of ACOS-GOSAT X_{CO_2} sampling on the inference of terrestrial biospheric sources and sinks. *Tellus B* **66**: 22486, doi: 10.3402/tellusb.v66.22486.

Müller J, Stavrakou T. 2005. Inversion of CO and NOx emissions using the adjoint of the IMAGES model. *Atmos. Chem. Phys.* **5**: 1157–1186.

Nassar R, Jones DBA, Suntharalingam P, Chen JM, Andres RJ, Wecht KJ, Yantosca RM, Kulawik SS, Bowman KW, Worden JR, Machida T, Matsueda H. 2010. Modeling global atmospheric CO_2 with improved emission inventories and CO_2 production from the oxidation of other carbon species. *Geosci. Model Dev.* **3**: 689–716.

Nassar R, Jones DBA, Kulawik SS, Worden JR, Bowman KW, Andres RJ, Suntharalingam P, Chen JM, Brenninkmeijer CAM, Schuck TJ, Conway TJ, Worthy DE. 2011. Inverse modeling of CO_2 sources and sinks using satellite observations of CO_2 from TES and surface flask measurements. *Atmos. Chem. Phys.* **11**: 6029–6047.

Nazareth L. 1979. Relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms. *SIAM J. Numer. Anal.* **16**: 794–800.

Nocedal J. 1980. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **35**: 773–782.

Nocedal J, Wright S. 2006. *Numerical Optimization* (2nd edn). Springer: Berlin.

O'Dell CW, Connor B, Boesch H, O'Brien D, Frankenberg C, Castano R, Christi M, Crisp D, Eldering A, Fisher B, Gunson M, McDuffie J, Miller CE, Natraj V, Oyafuso F, Polonsky I, Smyth M, Taylor T, Toon GC, Wennberg PO, Wunch D. 2012. The ACOS CO_2 retrieval algorithm –Part 1: Description and validation against synthetic observations. *Atmos. Meas. Tech.* **5**: 99–121.

Rabier F, Courtier P. 1992. Four-dimensional assimilation in the presence of baroclinic instability. *Q. J. R. Meteorol. Soc.* **118**: 649–672.

Sherman J, Morrison WJ. 1949. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Ann. Math. Statist.* **20**: 620–624, doi: 10.1214/aoms/1177729959.

Singh K, Jardak M, Sandu A, Bowman K, Lee M, Jones D. 2011. Construction of non-diagonal background-error covariance matrices for global chemical data assimilation. *Geosci. Model Dev.* **4**: 299–316.

Tarantola A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM: Philadelphia, PA, doi: 10.1137/1.9780898717921.

Tshimanga J, Gratton S, Weaver AT, Sartenaer A. 2008. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **134**: 751–769.

Veersé F, Auroux D, Fisher M. 2000. Limited-memory BFGS diagonal preconditioners for a data assimilation problem in meteorology. *Optim. Eng.* **1**: 323–339.

Yokota T, Yoshida Y, Eguchi N, Ota Y, Tanaka T, Watanabe H, Maksyutov S. 2009. Global concentrations of CO_2 and CH_4 retrieved from GOSAT: First preliminary results. *SOLA* **5**: 160–163.

