The Maximum Likelihood Ensemble Filter as a non-differentiable minimization algorithm

Milija Zupanski[1*], I. Michael Navon[2], and Dusanka Zupanski[1]

[1]*Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO*

[2] *School of Computational Science and Department of Mathematics, Florida State University, Tallahassee, FL*

Abstract: The Maximum Likelihood Ensemble Filter (MLEF) equations are derived without the differentiability requirement for the prediction model and for the observation operators. Derivation reveals that a new non-differentiable minimization method can be defined as a generalization of the gradient-based unconstrained methods, such as the preconditioned conjugate-gradient and quasi-Newton methods. In the new minimization algorithm the vector of first order increments of the cost function is defined as a generalized gradient, while the symmetric matrix of second order increments of the cost function is defined as a generalized Hessian matrix. In the case of differentiable observation operators, the minimization algorithm reduces to the standard gradient-based form.

The non-differentiable aspect of the MLEF algorithm is illustrated in an example with one-dimensional Burgers model and simulated observations. The MLEF algorithm has a robust performance, producing satisfactory results for tested non-differentiable observation operators.

---

[*] Correspondence to: Milija Zupanski, Cooperative Institute for Research in the Atmosphere, Colorado State University, 1375 Campus Delivery, Fort Collins, CO 80523-1375; E-mail: ZupanskiM@cira.colostate.edu

1   Introduction

The maximum likelihood ensemble filter (MLEF) is an ensemble data assimilation algorithm based on control theory (Zupanski 2005; Zupanski and Zupanski 2006). The MLEF is a posterior maximum likelihood approach, in a sense that it calculates the optimal state as the maximum of the probability density function (PDF), while most of the ensemble data assimilation methodologies used in meteorology and oceanography are based on the minimum variance approach (e.g., Evensen 1994; Houtekamer and Mitchell 1998; Bishop et al. 2001; Whitaker and Hamill 2002; Anderson 2003; Ott et al. 2004). The maximum of the PDF is found by an iterative minimization of the cost function derived from a multivariate posterior PDF. The iterative minimization is an important component of the MLEF since it provides practical means for finding the nonlinear analysis solution. The process of minimization produces both the most likely state and associated uncertainty.

The MLEF was successfully tested in applications with various weather prediction and related models, such as the Korteweg-de Vries-Burgers model (Zupanski 2005; Zupanski and Zupanski 2006), the Colorado State University (CSU) global shallow water model (Zupanski et al. 2006; Uzunoglu et al. 2007), the Large-Eddy Simulation (LES) model (Carrio et al. 2007), the National Aeronautics and Space Administration GEOS-5 column precipitation model (Zupanski et al. 2007b), and the CSU Lagrangian Particle Dispersion Model (LPDM) (Zupanski et al. 2007a). In all those applications a nonlinear conjugate-gradient method (e.g., Gill et al. 1981) was used for minimization of the cost function. As all other unconstrained gradient-based minimization algorithms, the nonlinear conjugate-gradient method requires the cost function to be at least twice differentiable. The first derivative of the cost function is required for the gradient, and the second derivative, or its approximation, is required for the Hessian preconditioning.

Unfortunately, the differentiability requirement is not necessarily satisfied in applications to realistic problems. In particular, the physical processes related to clouds and precipitation typically include non-differentiable operators. For example, it is known that cumulus convection parameterization in weather and climate introduces a significant

discontinuity in the first and higher-order derivatives (e.g., Verlinde and Cotton 1993; Zupanski 1993; Tsuyuki 1997; Xu and Gao 1999; Zhang et al. 2000). A similar discontinuity problem can be identified for observation operators as well. In satellite radiance assimilation, for example, a forward model for all-weather conditions is a non-differentiable observation operator. This follows from the fact that, depending on the value of the state vector (i.e. cloudy or clear), various forms of the forward operator will be chosen. For example, the cloud property model and the gas extinction model are only included in the presence of clouds, leading to different formulations of the forward operator in the presence of clouds and without clouds (Greenwald et al. 2002). Other non-differentiable operator examples can be found whenever a weather regime defined by the state vector defines different forms of the observation operator.

Common methods for solving non-differentiable (non-smooth) minimization are based on sub-differentials and bundle algorithms (Clarke 1983; Lemarechal and Zowe 1994; Nesterov 2005). Bundle algorithms were tested in optimal control problems of flow with discontinuities (Homescu and Navon 2003) using the PVAR software (Luksan and Vlcek 2001), and also in variatonal data assimilation (Zhang et al. 2000) using the bundle algorithm of Lemarechal (1977). An explicit knowledge of the minimization space (e.g. its basis or span-vectors), known in ensemble data assimilation, creates an opportunity to exploit alternative means for non-differentiable minimization without the need to define gradients, sub-gradients, or their approximations. Such an approach will be pursued here.

In this paper we address the differentiability requirement for the cost function by presenting an alternative derivation of the MLEF. For the first time the validity of the Taylor series expansion is not assumed, thus the differentiability of the cost function is not required. Since no limitation of using the first or second order Taylor formula approximation is imposed, the analysis and forecast ensemble perturbations are not restricted to be small. Under these relaxed conditions, the MLEF is formulated as a nonlinear filtering algorithm that allows non-differentiable models and observation operators.

An important consequence of this derivation is that the optimization algorithm used within the MLEF can be now viewed as a non-differentiable minimization

algorithm. For differentiable functions the minimization reduces to standard gradient-based algorithms, with an implicit Hessian preconditioning. In particular, the MLEF algorithm is presented as a non-differentiable generalization of the nonlinear conjugate-gradient and the BFGS quasi-Newton algorithms (Luenberger 1984; Gill et al. 1981). In order to illustrate the potential of the non-differentiable minimization used in the MLEF, a one-dimensional Burgers model simulating a shock wave is employed. Unlike in previous MLEF applications, we include a challenging non-differentiable observation operator with discontinuities in the function and in all its derivatives.

The paper is organized as follows. The new MLEF derivation under relaxed conditions is presented in Section 2. In Section 3 we describe the experimental design, model and observations. The results are presented in Section 4, and the conclusions are drawn in Section 5.

2   Non-differentiable MLEF formulation

Let the state space be denoted $\mathbb{S} \in \Re^{N_S}$, where N_S denotes its dimension, and let $\boldsymbol{x} \in \mathbb{S}$ be a state vector. We refer to the set of state vectors $\left\{ \boldsymbol{x}_i \in \mathbb{S} \; ; \; (i=1,\ldots,N_E) \right\}$ as ensembles, and to the space $\mathbb{E} \in \Re^{N_E}$ of dimension N_E as an ensemble space.

In order to begin ensemble data assimilation, the initial state vector and its uncertainty need to be specified. Let the initial state vector be denoted $\boldsymbol{x}^0$, and let the initial $N_S \times N_E$ square-root error covariance be denoted $\boldsymbol{P}_0^{1/2} : \mathbb{E} \to \mathbb{S}$ with columns $\left\{ \boldsymbol{p}_i^0 \in \mathbb{S} ; (i=1,\ldots,N_E) \right\}$. The initial state vector and the initial square-root error covariance define a set of initial conditions

$$\boldsymbol{x}_i^0 = \boldsymbol{x}^0 + \boldsymbol{p}_i^0 ; \qquad (i=1,\ldots,N_E) \tag{1}$$

2.1. Prediction

The predictive step of the MLEF (and any other filter) addresses the means of transporting the uncertainty span-vectors from the current analysis time to the next

analysis time. A nonlinear dynamical model $M : \mathbb{S} \to \mathbb{S}$ transports the state vector according to

$$\boldsymbol{x}_t = M(\boldsymbol{x}_{t\text{-}1}) \tag{2}$$

where t-1 and t refer to the current and the next analysis times, respectively. Note that the model error is neglected in Equation (2) to simplify the derivation. In order to keep the notation manageable, we omit the time index in the remainder of the paper, unless suggested otherwise. The *forecast* increment resulting from the i-th analysis increment is

$$\Delta_i \boldsymbol{x}^f = \boldsymbol{x}_i^f - \boldsymbol{x}^f = M(\boldsymbol{x}_i^a) - M(\boldsymbol{x}^a) = M(\boldsymbol{x}^a + \boldsymbol{p}_i^a) - M(\boldsymbol{x}^a) \ . \tag{3}$$

where the superscripts a and f refer to analysis and forecast, respectively. The vectors $\{\boldsymbol{p}_i^a \in \mathbb{S}; (i = 1, \dots, N_E)\}$ represent the columns of the square-root analysis error covariance. After defining $\boldsymbol{p}_i^f = \Delta_i \boldsymbol{x}^f$, the square-root forecast error covariance is

$$\boldsymbol{P}_f^{1/2} = \begin{bmatrix} \boldsymbol{p}_1^f & \boldsymbol{p}_2^f & \cdots & \boldsymbol{p}_{N_E}^f \end{bmatrix} \qquad \boldsymbol{p}_i^f = M(\boldsymbol{x}^a + \boldsymbol{p}_i^a) - M(\boldsymbol{x}^a) \tag{4}$$

where $\boldsymbol{P}_f^{1/2} : \mathbb{E} \to \mathbb{S}$ is a $N_S \times N_E$ matrix with columns $\{\boldsymbol{p}_i^f \in \mathbb{S}; (i = 1, \dots, N_E)\}$.

The Equation (3) represents the transport of uncertainty span-vectors in time by nonlinear model dynamics. The MLEF forecast step allows the nonlinear model operator and large analysis increments to be included without typical restrictions, such as linearity and differentiability. For small analysis increments, however, the forecast error covariance formulation (4) reveals that the forecast step of the MLEF is closely related to the Kalman filters (e.g., Jazwinski 1970), and to the SEEK filter (Pham et al. 1998; Rozier et al. 2007).

2.2. Analysis

The analysis is corrected in the subspace defined by the forecast error covariance matrix (Jazwinski 1970). Using the Equation (4) one can define the analysis correction subspace

$$\mathbb{A} = span\left\{ \boldsymbol{p}_1^f, \boldsymbol{p}_2^f, \dots, \boldsymbol{p}_{N_E}^f \right\} \qquad \mathbb{A} \subseteq \mathbb{S}. \tag{5}$$

Then, an arbitrary vector $\boldsymbol{x} - \boldsymbol{x}^f \in \mathbb{A}$ can be expressed as a linear combination

$$\boldsymbol{x} - \boldsymbol{x}^f = w_1 \boldsymbol{p}_1^f + w_2 \boldsymbol{p}_2^f + \cdots + w_{N_E} \boldsymbol{p}_{N_E}^f = \boldsymbol{P}_f^{1/2} \boldsymbol{w} \qquad\qquad \boldsymbol{w} = \left(w_1, w_2, \ldots, w_{N_E} \right)^T \in \mathbb{E}. \quad (6)$$

The transformation (6) links the analysis correction subspace with the ensemble space, and shows that $\boldsymbol{P}_f^{1/2} : \mathbb{E} \to \mathbb{A}$.

Until now, the differentiability of the dynamical model M was not required, and no specific assumption about the probability distribution of the analysis or forecast increments was necessary. In the analysis, however, some assumptions will be required. Often assumed, as done here, is that probability distribution of the initial conditions errors and the observation errors are Gaussian. The novelty is that a commonly used differentiability assumption will be relaxed, thus a more general formulation of the MLEF analysis solution will be derived. Note that it is possible to relax the Gaussian assumption and develop non-Gaussian data assimilation framework (e.g., van Leeuwen 2003; Abramov and Majda 2004; Majda et al. 2004; Fletcher and Zupanski 2006a,b), but this will not be pursued here in order to simplify the presentation.

2.3.1. Cost function

In the MLEF, the optimal set of coefficients $\{ w_i : \ i = 1, \ldots, N_E \}$ is obtained by maximizing the posterior conditional probability. In practice this is achieved by an iterative minimization of a cost function (e.g., Lorenc 1986)

$$J(\boldsymbol{x}) = \frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{x}^f \right)^T \boldsymbol{P}_f^{-1} \left(\boldsymbol{x} - \boldsymbol{x}^f \right) + \frac{1}{2} \left[\boldsymbol{y} - H(\boldsymbol{x}) \right]^T \boldsymbol{R}^{-1} \left[\boldsymbol{y} - H(\boldsymbol{x}) \right] \qquad\qquad (7)$$

where $\boldsymbol{R} : \mathbb{O} \to \mathbb{O}$ is the observation error covariance, $\mathbb{O} \in \mathfrak{R}^{N_O}$ is the observation space, N_O is the dimension of $\mathbb{O}$, $\boldsymbol{y} \in \mathbb{O}$ is the observation vector, and $H : \mathbb{S} \to \mathbb{O}$ is a nonlinear and/or non-differentiable observation operator. Since the matrix $\boldsymbol{P}_f$ is defined using ensemble forecast increments, the minimization of the cost function will involve a search in the analysis correction subspace $\mathbb{A}$.

Let consider an increment of the cost function, $\Delta J(\boldsymbol{x}) = J(\boldsymbol{x} + \Delta \boldsymbol{x}) - J(\boldsymbol{x})$, for $\Delta \boldsymbol{x} \in \mathbb{A}$. In principle, the minimization of $\Delta J(\boldsymbol{x})$ is equivalent to minimizing $J(\boldsymbol{x})$ (e.g., Luenberger 1984). Direct substitution of $\boldsymbol{x} + \Delta \boldsymbol{x}$ in the Equation (7) results in

$$J(x + \Delta x) = J(x) + (\Delta x)^T P_f^{-1}(x - x^f) - [H(x + \Delta x) - H(x)]^T R^{-1}[y - H(x)] +$$
$$+ \frac{1}{2}(\Delta x)^T P_f^{-1}(\Delta x) + \frac{1}{2}[H(x + \Delta x) - H(x)]^T R^{-1}[H(x + \Delta x) - H(x)] \quad . \tag{8}$$

Note that for *differentiable* operator *H*, the expansion (8) reduces to

$$J(x + \Delta x) = J(x) + (\Delta x)^T P_f^{-1}(x - x^f) - (\Delta x)^T \left[\frac{\partial H}{\partial x}\right]^T R^{-1}[y - H(x)] +$$
$$+ \frac{1}{2}(\Delta x)^T P_f^{-1}(\Delta x) + \frac{1}{2}(\Delta x)^T \left[\frac{\partial H}{\partial x}\right]^T R^{-1}\left[\frac{\partial H}{\partial x}\right](\Delta x) + O(\|\Delta x\|^3) \tag{9}$$

which is equivalent to a second-order Taylor series expansion of $J(x)$ in the vicinity of x. One can note that the Taylor expansion (9) has a remainder $O(\|\Delta x\|^3)$, due to neglecting the higher-order nonlinear terms (where $\|.\|$ denotes a norm). On the other hand, the expansion (8) does not have a remainder since the use of total increments accounts for all higher-order nonlinear terms. Therefore, the formula (8) may be viewed as a generalization of the Taylor expansion of *J*. This apparent similarity can be used to define a generalization of the gradient vector and the Hessian matrix that could be used in minimization, to include nonlinear and non-differentiable operator *H*.

Since the analysis correction subspace $\mathbb{A}$ is already defined (Equation (5)), the increments Δx in the direction of known span-vectors $\Delta_i x^f = \{p_i^f \in \mathbb{A}, i = 1,...,N_E\}$ are considered. By comparing Equations (8) and (9), one can identify the *i*-th component of a generalized first derivative of *J*, denoted $\nabla_i J$,

$$\nabla_i J(x) = (\Delta_i x^f)^T P_f^{-1}(x - x^f) - [H(x + \Delta_i x^f) - H(x)]^T R^{-1}[y - H(x)] \tag{10}$$

and the (*i,j*)-th element of a generalized second derivative of *J*, denoted $\nabla_{i,j}^2 J$,

$$\nabla_{i,j}^2 J(x) = (\Delta_i x^f)^T P_f^{-1}(\Delta_j x^f) + [H(x + \Delta_i x^f) - H(x)]^T R^{-1}[H(x + \Delta_j x^f) - H(x)] \tag{11}$$

where $\{\nabla_{i,j}^2 = \nabla_i(\nabla_j) : i, j \leq N_E\}$. If we define an $N_O \times N_E$ observation perturbation matrix, $Z : \mathbb{S} \rightarrow \mathbb{O}$

$$Z(x) = \begin{bmatrix} z_1(x) & z_2(x) & \cdot & \cdot & z_{N_E}(x) \end{bmatrix} \qquad z_i(x) = R^{-1/2}[H(x + p_i^f) - H(x)], \tag{12}$$

the *generalized* first and second derivatives are

$$\nabla_G J(\boldsymbol{x}) = \boldsymbol{P}_f^{-1/2}\left(\boldsymbol{x} - \boldsymbol{x}^f\right) - \left(\boldsymbol{Z}(\boldsymbol{x})\right)^T \boldsymbol{R}^{-1/2}\left[\boldsymbol{y} - H(\boldsymbol{x})\right] \tag{13}$$

$$\nabla_G^2 J(\boldsymbol{x}) = \boldsymbol{I} + \left(\boldsymbol{Z}(\boldsymbol{x})\right)^T \boldsymbol{Z}(\boldsymbol{x}) \tag{14}$$

Note that the generalized first derivative is a N_E-dimensional vector, and the generalized second derivative is a $N_E \times N_E$ matrix, i.e. both are defined in ensemble space $\mathbb{E}$. The Equations (13) and (14) are not approximations to the true derivatives since all nonlinear terms are included in the matrix $\boldsymbol{Z}(\boldsymbol{x})$. In absence of better terminology, the term "derivative" is used only to indicate that for differentiable cost function and for small perturbations $\boldsymbol{p}_i^f$ the Equations (13) and (14) would reduce to finite-difference approximations of directional derivatives.

The similarity of the generalized gradient (Equation (13)) with the generalized gradient in the subgradient method (e.g. Zhang et al. 2000), $\left[\partial J(\boldsymbol{x})\right]^T \Delta_i \boldsymbol{x}^f$, reveals that the formulation adopted here does satisfy the requirement for the subgradient, i.e. $J(\boldsymbol{x} + \Delta_i \boldsymbol{x}^f) \geq J(\boldsymbol{x}) + \left[\nabla_G J(\boldsymbol{x})\right]_i$ since $\nabla_G^2 J(\boldsymbol{x}) \geq 0$ for all $\Delta_i \boldsymbol{x}^f \in \mathbb{A}$. However, in our formulation the increments $\Delta_i \boldsymbol{x}^f$ are included in the nonlinear perturbation of the observation operator through Equation (12), and thus cannot be separated into the subgradient and the perturbation. In other words, our method includes the minimization space span (or basis) vectors as inseparable components of the generalized derivatives definition.

Since the MLEF formulation employs finite differences (i.e. increments), it is interesting to compare the MLEF derivatives (Eqs. (13) and (14)) with the finite-difference approximations of derivatives. The finite-difference approximations to directional derivatives of the cost function (7) are $\left[\partial J / \partial x_i\right]\Delta_i \boldsymbol{x} \approx J(\boldsymbol{x} + \Delta_i \boldsymbol{x}) - J(\boldsymbol{x})$ and $\left[\partial^2 J / \partial x_i \partial x_j\right]\Delta_i \boldsymbol{x} \Delta_j \boldsymbol{x} = J(\boldsymbol{x} + \Delta_i \boldsymbol{x} + \Delta_j \boldsymbol{x}) - J(\boldsymbol{x} + \Delta_i \boldsymbol{x}) - J(\boldsymbol{x} + \Delta_j \boldsymbol{x}) + J(\boldsymbol{x})$. Note that the finite-difference representation of the first derivative already includes all terms of Eq.(8), i.e. includes both the first and the second derivatives used by the MLEF. The finite-difference approximation of the second derivative is also not resembling the MLEF formulation (e.g. Eq.(14)). Therefore, the MLEF is not a finite-difference approximation

to derivatives. The finite-difference approximation to derivatives has some advantages related to its simplicity, but it may require a control of positive-definiteness of the Hessian, and possibly additional computational time due to the term $J(x + \Delta_i x + \Delta_j x)$ in the Hessian.

2.3.2. Generalized Hessian preconditioning

A common starting point of minimization is the state vector $x = x^f$ (corresponding to w=0 in Equation (6)), since it represents the best knowledge of the dynamical state prior to taking into account the observations. Since the optimal preconditioning is defined as an inverse square-root Hessian matrix (e.g., Axelsson and Barker 1984), one can utilize Equation (14) to define Hessian preconditioning as a change of variable

$$w = \left[\nabla_G^2 J(x^f) \right]^{-1/2} \zeta = \left[I + \left(Z(x^f) \right)^T Z(x^f) \right]^{-1/2} \zeta \tag{15}$$

where $\zeta \in \mathbb{E}$ is the control vector of dimension N_E, and $Z(x^f)$ is obtained by substituting $x = x^f$ in Equation (12). As explained in Zupanski (2005), and equivalent to the procedure used in the ensemble transform Kalman filter (ETKF, Bishop et al. 2001), one can perform an eigenvalue decomposition of $\nabla_G^2 J(x^f)$ to obtain $\left[\nabla_G^2 J(x^f) \right]^{-1/2} = V(I + \Lambda)^{-1/2} V^T$, where V is the eigenvector matrix and Λ is the eigenvalue matrix of $\nabla_G^2 J(x^f)$. Note that the MLEF transformation calculates a symmetric square-root matrix, corresponding to the ETKF transform with simplex improvement (Wang et al. 2004; Wei et al. 2006).

By combining Equations (6) and (15), one obtains the *generalized Hessian preconditioning* in state space, in the form of the change of variable

$$x - x^f = G^{1/2} \zeta \qquad\qquad G^{1/2} = P_f^{1/2} \left[I + \left(Z(x^f) \right)^T Z(x^f) \right]^{-1/2}. \tag{16}$$

The matrix $G^{1/2} : \mathbb{E} \to \mathbb{S}$ is a $N_S \times N_E$ matrix, and it represents the inverse of the square-root generalized Hessian matrix estimated at the initial point of minimization.

Once the Hessian preconditioning is accomplished, one can begin with calculation of preconditioned generalized gradients. An iterative minimization produces ζ_k at iteration k according to $\zeta_k = \zeta_{k-1} + \alpha_{k-1} d_{k-1}$, where $\alpha_{k-1} \in \Re^1$ and $d_{k-1} \in \mathbb{E}$ are the step-length and the descent direction at the k-1-th iteration, respectively. Using the change of variable (16), the state vector at k-th minimization iteration is related to the control vector as

$$x_k = x^f + G^{1/2} \zeta_k \tag{17}$$

The *preconditioned generalized gradient* at the k-th minimization iteration is obtained by employing the Hessian preconditioning formulation (16) and evaluating (13) at x_k

$$\nabla_G J(x_k) = \left[I + \left(Z(x^f) \right)^T Z(x^f) \right]^{-1} \zeta_k - \left(Z(x_k) \right)^T R^{-1/2} \left[y - H(x_k) \right] \tag{18}$$

where $Z(x_k)$ is obtained by substituting $x = x_k$ in Equation (12).

2.3.3. Analysis error covariance

In order to complete the non-differentiable formulation of the MLEF, an analysis (e.g., posterior) error covariance matrix, quantifying the uncertainties of the analysis, is required. The equivalence between the inverse Hessian at the optimal point and the posterior error covariance (e.g., Fisher and Courtier 1995; Veerse 1999) is exploited in the MLEF algorithm. More detailed examination of the relation between the inverse Hessian and analysis error covariance in nonlinear problems can be found in Gejadze et al. (2007). Since the generalized Hessian in ensemble space is given by Equation (16), the analysis (posterior) error covariance in ensemble space is defined as

$$(P_w)_a = \left(\nabla_G^2 J(x^a) \right)^{-1} = \left(I + \left(Z(x^a) \right)^T Z(x^a) \right)^{-1} \tag{19}$$

where $Z(x^a)$ is obtained by substituting $x = x^a$ in Equation (12). The analysis error covariance in state space can be obtained by utilizing the change of variable (6) to define the *true* and *optimal* state vectors, x^t and x^a, as $x^t - x^f = P_f^{1/2} w^t$ and $x^a - x^f = P_f^{1/2} w^a$, where w^t and w^a are the *true* and *optimal* control vectors in ensemble space,

respectively. Then, the error of state vector is related to the error of the control vector in ensemble space according to

$$x^a - x^t = P_f^{1/2}(w^a - w^t) \ .$$ (20)

By taking the mathematical expectation of an outer product of (20), and utilizing (19), one obtains the analysis error covariance in state space

$$P_a = P_f^{1/2}\left[I + \left(Z(x^a)\right)^T Z(x^a)\right]^{-1} P_f^{T/2}$$ (21)

As suggested in section 2.2, the columns of the square-root analysis error covariance, denoted $P_a^{1/2}$, are used to define the initial perturbations for the ensemble forecast. Then, the matrix $P_a^{1/2}$ can be written in a column form as

$$P_a^{1/2} = \begin{bmatrix} p_1^a & p_2^a & \cdots & p_{N_E}^a \end{bmatrix} \qquad p_i^a = \left(P_f^{1/2}\left[I + \left(Z(x^a)\right)^T Z(x^a)\right]^{-1/2} \right)_i$$ (22)

The matrix $P_a^{1/2}$ is a $N_S \times N_E$ matrix. In principle, instead of the relation (19) for the inverse generalized Hessian in ensemble space, one could use the BFGS inverse Hessian update (e.g., Veerse 1999), or some other estimate of the inverse Hessian at the optimal point (e.g., Gejadze et al. 2007). The expression (19) is currently being used in the MLEF algorithm.

3   Non-differentiable minimization algorithms

In this section, two non-differentiable minimization algorithms generalized using the derivation from Section 2 will be formulated. The first algorithm is the generalized nonlinear conjugate-gradient minimization algorithm, presented for both the Fletcher-Reeves and the Polak-Ribiere formulations (e.g., Luenberger 1984).

Algorithm 1. (Generalized nonlinear conjugate-gradient)

Choose starting point $\mathbf{x}_0 = \mathbf{x}^f \Leftrightarrow \zeta_0 = 0$ and define $\mathbf{d}_0 = -\left(\nabla_G J\right)_0 = -\nabla_G J(\mathbf{x}_0)$;

$k \leftarrow 0$;

repeat

- Compute $\left(\nabla_G J\right)_{k+1} = \nabla_G J(\mathbf{x}_{k+1})$

- Set $\mathbf{d}_{k+1} = -\left(\nabla_G J\right)_{k+1} + \beta_k \mathbf{d}_k$

where $\beta_k = \dfrac{\left(\nabla_G J\right)_{k+1}^T \left(\nabla_G J\right)_{k+1}}{\left(\nabla_G J\right)_k^T \left(\nabla_G J\right)_k}$ (Fletcher-Reeves), or

$$\beta_k = \dfrac{\left[\left(\nabla_G J\right)_{k+1} - \left(\nabla_G J\right)_k\right]^T \left(\nabla_G J\right)_{k+1}}{\left(\nabla_G J\right)_k^T \left(\nabla_G J\right)_k} \quad \text{(Polak-Ribiere)}$$

- Update $\zeta_{k+1} = \zeta_k + \alpha_k \mathbf{d}_k$ and $x_{k+1} = x^f + \mathbf{G}^{1/2}\zeta_{k+1}$ where α_k minimizes $J(x^f + \mathbf{G}^{1/2}\zeta_{k+1})$
until convergence.

The second minimization algorithm is the generalized BFGS quasi-Newton algorithm developed from the differentiable form (Nocedal 1980; Liu and Nocedal 1989). The limited-memory formulation is a straightforward extension, obtained by discarding some terms in the inverse Hessian (e.g., Nocedal and Wright 1999).

Algorithm 2. (Generalized quasi-Newton)

Choose starting point $\mathbf{x}_0 = \mathbf{x}^f \Leftrightarrow \zeta_0 = 0$ and define $\mathbf{d}_0 = -\left(\nabla_G J\right)_0 = -\nabla_G J(\mathbf{x}_0)$;

$k \leftarrow 0$;

repeat

- Compute $\left(\nabla_G J\right)_{k+1} = \nabla_G J(\mathbf{x}_{k+1})$

- Set $\mathbf{s}_k = \zeta_{k+1} - \zeta_k$ and $\mathbf{y}_k = \left(\nabla_G J\right)_{k+1} - \left(\nabla_G J\right)_k$

- Set $\rho_k = 1 / \mathbf{y}_k^T \mathbf{s}_k$ and $\upsilon_k = \left(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T\right)$

- Compute $\mathbf{H}_{k+1} = \upsilon_k^T \upsilon_{k-1}^T \cdots \upsilon_0^T \upsilon_0 \cdots \upsilon_{k-1} \upsilon_k + \upsilon_k^T \upsilon_{k-1}^T \cdots \upsilon_1^T \rho_0 \mathbf{s}_0 \mathbf{s}_0^T \upsilon_1 \cdots \upsilon_{k-1} \upsilon_k + \cdots$

$$+ \upsilon_k^T \rho_{k-1} \mathbf{s}_{k-1} \mathbf{s}_{k-1}^T \upsilon_k + \rho_k \mathbf{s}_k \mathbf{s}_k^T$$

- Set $\mathbf{d}_{k+1} = -\mathbf{H}_{k+1}\left(\nabla_G J\right)_{k+1}$

- Update $\zeta_{k+1} = \zeta_k + \alpha_k \mathbf{d}_k$ and $x_{k+1} = x^f + \mathbf{G}^{1/2}\zeta_{k+1}$ where α_k minimizes $J(x^f + \mathbf{G}^{1/2}\zeta_{k+1})$
until convergence.

The above minimization algorithms show that the MLEF could be used as a stand-alone non-differentiable minimization algorithm in applications other than ensemble data assimilation. In principle, there are two possible means to perform minimization using the MLEF: (*i*) if relevant directions are known, the MLEF can be used as a reduced-rank minimization algorithm in the subspace spanned by relevant directions, and (*ii*) if relevant directions are not known, one can define a basis representing the full space and use it to perform a regular, full-rank minimization with MLEF. If there are means to define the set of relevant directions, computational savings due to the reduced-rank formulation would make this option advantageous. On the other hand, the full-rank option (*ii*) is a straightforward extension of the standard conjugate-gradient and quasi-Newton algorithms, thus it may be easier to apply in principle. In this paper we chose the option (*i*), since the columns of the square-root forecast error covariance represent relevant directions.

In the above formulations we did not specify the line-search algorithm. Although this is an important aspect of non-differentiable minimization, in our current implementation a simple line-search method is used (e.g., Derber 1989), which involves one function evaluation per minimization iteration. The same line-search method was used in the GRAD and the MLEF experiments.

The computational cost of the MLEF minimization is $N_E + 2$ function evaluations per minimization iteration. This estimate includes the gradient and the Hessian calculations. The computational cost increases almost linearly with the number of ensembles. As for other ensemble data assimilation algorithms, parallel processing can significantly increase the speed of the MLEF minimization since the communication between processors is almost negligible.

On the other hand, the computational cost of the gradient-based minimization depends mostly on the cost of the gradient calculation. In our implementation of the gradient-based method the cost per iteration is $2N_E + 2$, i.e. about two times more than the MLEF cost. In meteorological applications the adjoint model (e.g., LeDimet and Talagrand 1986) is typically used to compute the gradient. Its cost per iteration is about

2-5 function evaluations, depending on the complexity of the operator. This means that the cost per iteration minimization is 4-7 function evaluations. In typical operational meteorological applications (without the Hessian calculation) there are about 60-80 minimization iterations, making the cost of minimization about 300-500 function evaluations. For similar complexity of the problem the MLEF minimization requires only one minimization iteration. This approximate calculation makes the cost of the gradient-based and MLEF minimization similar for the ensemble size of 300-500. With Hessian calculation included, the number of iterations would be smaller, but the total cost would likely increase. One should keep in mind, however, that the actual cost would strongly depend on the complexity of the problem and on the computational capabilities.

4   Experimental design and results

Two nonlinear observation operators will be tested, a quadratic and a cubic operator. In addition, each of the operators will have a differentiable and a non-differentiable form. The differentiable observation operators are defined as

(a) $\qquad H(x) = x^2$ $\qquad\qquad\qquad\qquad (b)$ $\qquad H(x) = x^3$ $\qquad\qquad\qquad\qquad (23)$

while the corresponding non-differentiable observation operators are defined as

(a) $\qquad H(x) = \begin{cases} x^2 & \text{for } x \geq 0.5 \\ -x^2 & \text{for } x < 0.5 \end{cases}$ $\qquad (b)$ $\qquad H(x) = \begin{cases} x^3 & \text{for } x \geq 0.5 \\ -x^3 & \text{for } x < 0.5 \end{cases}$ $\qquad (24)$

The non-differentiable operators given by Equation (24) are shown in Figure 1. Although they may appear relatively simple, both observation operators do have discontinuities in the function and its derivatives. It is interesting to note that the quadratic non-differentiable operator has a more pronounced discontinuity than the cubic operator. This makes the use of quadratic non-differentiable operator more challenging for minimization. The opposite is expected for differentiable operators, since the quadratic operator is less nonlinear than the cubic operator.

The prediction model we use is the one-dimensional Burgers equation (Burgers 1948)

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = v\frac{\partial^2 u}{\partial x^2} \qquad\qquad (25)$$

where v is a viscosity coefficient and u is the velocity. This equation is often used in fluid dynamics for simulation of nonlinear waves, shock formation, and turbulence. The Equation (25) is solved numerically using centered space differences and Lax-Wendroff time integrating scheme (Fletcher 1991). In this paper we use the Burgers equation to simulate a propagating shock wave (Akella and Navon 2006). The dimension of the state vector is 81.

We conduct a twin model experiment, in which the prediction from one set of initial conditions is defined as 'truth' (denoted TRUE), while the prediction from a different set of initial conditions is defined as 'experimental' (denoted EXP). In our case the EXP initial conditions are defined as a 40-time step old TRUE forecast. The TRUE and EXP initial conditions are shown in Figure 2. The figure indicates that the velocity values are typically between 0 and 1, and that the forecast lag is about 20 grid-points. Also, one can notice a steep velocity gradient that is simulating a shock wave. The observations are created by adding Gaussian random perturbation to the TRUE model first guess, $H(x^{true})$, with zero mean and standard deviations of 8.0×10^{-2} (when using the quadratic observation operator), and 7.0×10^{-4} (when using the cubic observation operator). Random perturbations are added at each grid-point, implying that there are 81 observations. The time frequency of observations is 20 model time steps, which also defines the length of data assimilation cycle. We create observations during 20 data assimilation cycles, but most relevant adjustments happen during the first 5-10 data assimilation cycles.

All data assimilation experiments are done with 4 ensemble members, without error covariance localization and/or inflation. The initial ensemble perturbations are defined using lagged (time-shifted) forecasts that correspond to the EXP model run, centered about the initial time of the data assimilation cycle No.1. This approach employs

the so-called ergodic hypothesis, in which the time differences are used to represent the spatial differences. This initial set-up creates dynamically balanced perturbations, thus less noisy initial error covariance. In all other data assimilation cycles the ensemble perturbations are updated using Equation (22).

In order to test the non-differentiable and/or nonlinear minimization performance, for each of the two observation operators we conduct two data assimilation experiments using the nonlinear conjugate-gradient algorithm described in previous section: (*i*) with generalized Hessian preconditioning and generalized gradient (Equations (16) and (18), respectively), and (*ii*) with regular derivatives. The first experiment is denoted MLEF, since this is the standard form of the MLEF algorithm, and the second experiment is denoted GRAD to reflect its gradient-based characteristics. The regular derivatives are obtained by employing a linear approximation $H(\boldsymbol{x} + \boldsymbol{p}_i^f)\text{-}H(\boldsymbol{x}) \approx \left(\dfrac{\partial H}{\partial \boldsymbol{x}}\right)\boldsymbol{p}_i^f$ in the definition of the observation perturbation matrix (Equation (12)), i.e. by using $\boldsymbol{z}_i(\boldsymbol{x}) = \boldsymbol{R}^{-1/2}\left(\dfrac{\partial H}{\partial \boldsymbol{x}}\right)\boldsymbol{p}_i^f$ in Equations (16) and (18).

The difference between the observation operator gradients in the MLEF and the GRAD formulations comes from the higher-order terms in Taylor expansion, assuming that the function $H(x)$ is differentiable. Let consider the cubic observation operator (Eq.(23b)) as an example. By direct substitution, one can see that the difference between the MLEF and the GRAD formulations is $3x(\boldsymbol{p}_i^f)^2 + (\boldsymbol{p}_i^f)^3$. For nonlinear problems, the perturbation $\boldsymbol{p}_i^f$ can be large, thus making the difference between the gradients large. The difference becomes even more relevant for non-differentiable operators near the discontinuous point. Let consider an example with x=0.4 and $\boldsymbol{p}_i^f = 1.0$ (such that the perturbed operator crosses the discontinuity), again using the cubic observation operator as an example. Following from the above considerations and from the Eq.(24b), the MLEF gradient at point x=0.4 is equal to $H(\boldsymbol{x} + \boldsymbol{p}_i^f)\text{-}H(\boldsymbol{x})=(0.4+1.0)^3\text{-}(0.4)^3 = 2.68$,

while the GRAD gradient is equal to $\left(\dfrac{\partial H}{\partial \boldsymbol{x}}\right)\boldsymbol{p}_i^f = -3x^2\boldsymbol{p}_i^f = -(0.4)^2 \times 1.0 = -0.16$. This

implies that the MLEF and the GRAD would have gradients in the opposite directions!

The experimental results will indicate which of the two formulations has an advantage. Before we show the results, however, it should be noted that for the cubic observation operator the maximum allowed size of the control variable adjustment had to be restricted in the GRAD experiment, in order to prevent the minimization divergence. In the MLEF minimization this was not necessary, presumably because the algorithm is not relying on small perturbations of the Taylor expansion.

4.1 Non-differentiable observation operator

The results of the two algorithms over all 20 data assimilation cycles, using non-differentiable observation operators (Eq. (24)), are presented in terms of the analysis root-mean-squared (*RMS*) errors in Fig. 3. Knowing the true state, $\boldsymbol{x}^{true}$, the analysis *RMS* error is calculated as

$$RMS = \sqrt{\frac{1}{N_S}\sum_{i=1}^{N_S}\left(x_i^a - x_i^{true}\right)^2} \qquad (26)$$

For both observation operators the analysis RMS error in the MLEF experiment is smaller than in the GRAD experiment. For the quadratic observation operator this advantage is less obvious than for the cubic observation operator. This could be related to a more pronounced discontinuity noted for the quadratic non-differentiable observation operator (e.g., Fig.1). For the cubic observation operator, the MLEF analysis errors are smaller than GRAD errors 2-3 times during the first several cycles, eventually reaching 7-times smaller value in later cycles. However, all experiments eventually achieve negligible errors when the shock wave exits the right boundary of the integration domain.

An example of the difficulties of the GRAD estimation of the gradient is illustrated in Fig.4. The directional gradients for the cubic non-differentiable operator and for the ensemble member 2 are shown in the first minimization iteration of the first data assimilation cycle. A vertical dashed line, separating the two regions of the velocity U,

indicates the discontinuous point. In the region where U<0.5, the GRAD creates a gradient of the opposite sign to the MLEF gradient, as suggested in our discussion at the beginning of this section. In the region where U>0.5, gradients often change the sign, sometimes having an opposite sign as well, and the GRAD is generally of larger magnitude than the MLEF gradient. The linear approximation used in GRAD creates an apparent disadvantage of the gradient-based method, as implied from the larger analysis RMS errors (Fig.3).

   More details of the performance can be seen from Fig. 5, which shows the velocity analysis differences between the MLEF and GRAD experiments in the cubic non-differentiable observation operator experiment. Only first four data assimilation cycles are shown, since most important velocity adjustment occurs during these cycles. One can see that the analysis errors are systematically smaller in the MLEF experiment. It is also interesting to note that analysis errors are becoming more localized as new observations are assimilated, indicating the self-localizing characteristic of the MLEF. By the cycle 4, the MLEF achieves five times smaller maximum analysis error than the GRAD experiment. Since the only difference between the two experiments is in the minimization procedure, this indicates a superior performance of the MLEF non-differentiable minimization.

   In order to further examine the impact of the MLEF non-differentiable minimization algorithm, the cost function and the gradient norm in first data assimilation cycle are shown in Figs. 6 and 7, for quadratic and for cubic non-differentiable operators, respectively. The gradient norm is defined as $\|\boldsymbol{g}\| = \left(\boldsymbol{g}^T \boldsymbol{g}\right)^{1/2}$, where $\boldsymbol{g}$ denotes the gradient. Note that the gradient norm refers to the generalized gradient norm in the case of the MLEF. The results are shown for the first 20 minimization iterations, since the later iterations did not bring any relevant change.

   For quadratic operator, the cost functions in the MLEF and the GRAD experiments become almost the same (Fig.6a). The gradient norm (Fig.6b), however, shows a better convergence of the MLEF minimization, with gradient norm decreasing by several orders of magnitude. One can note the irregular behavior of the cost function,

with several jumps, as well as of the gradient norm. One can see that the cost function jumps match with the gradient norm jumps, suggesting that GRAD minimization has difficulties due to the gradient estimation. For cubic operator, the GRAD cost function decreased by one order of magnitude, while the MLEF cost function decreased by more than three orders of magnitude (Fig.7a). The gradient norm indicates a serious problem in the GRAD minimization, without an obvious reduction, while in the MLEF minimization the gradient norm was reduced by almost five orders of magnitude (Fig.7b).

4.2. Differentiable observation operator

One would expect that both experiments, especially the GRAD minimization, would perform better if non-differentiability of the observation operator were removed. In order to test this assumption, and to further examine the differences between the MLEF and GRAD minimization algorithms, we repeated similar experiments as in the section 4.1, except using the cubic differentiable observation operator given by Eq. (23b). As in section 4.1, we concentrate on minimization performance in the first data assimilation cycle. The results of the MLEF and GRAD minimization algorithms are shown in Fig. 8, in terms of the cost function and the gradient norm. One can see somewhat similar differences as in the Fig. 7, again indicating a superior MLEF minimization performance. A closer inspection of the cost functions in Figs. 7a and 8a shows that indeed differentiability did help. The MLEF minimization did produce faster convergence than in the non-differentiable operator case, and the shape of the cost function decrease indicates a well-behaved minimization without the difficulties observed in the cycles 2-9 of Fig. 7a. The cost function becomes flat after only 3-4 iterations, while in the non-differentiable case 9-10 iterations were necessary. A careful inspection of Figs. 7 and 8 indicates that the GRAD cost function shows a better performance compared to the non-differentiable case, mostly by continuing to decrease throughout the iterations, rather than reaching saturation as in the non-differentiable case. The gradient norms are also showing some improvement compared to the non-differentiable case. In the GRAD minimization, there is a slight overall decrease of the gradient norm, compared with the gradient norm increase during iterations 10-16 in the non-differentiable case. In the MLEF

minimization, there is also a larger decrease of the gradient norm, almost one more order of magnitude.

5   **Summary and Conclusions**

A new derivation of the MLEF algorithm is presented. It is shown that the same final equations as in the original formulation can be obtained without assuming differentiability and linearity of the prediction model and observation operators, as it would be typically done using the Taylor expansion. In order to generalize the nonlinear conjugate-gradient and quasi-Newton minimization algorithms we introduced a generalized gradient and generalized Hessian as non-differentiable equivalents of the standard gradient and the Hessian. For linear and differentiable operator H, the generalized gradient and Hessian formulations reduce to directional first and second derivatives in the direction of $\Delta(x_i^f)$ (e.g., Gill et al. 1981).

An implicit inclusion of higher-order nonlinear terms in the non-differentiable MLEF algorithm is important for nonlinear observation operators, being more accurate, but also by allowing larger perturbations to be included in minimization. Therefore, the MLEF system has a potential to work with challenging prediction models and observation operators encountered in geophysical and related applications, in principle non-differentiable and/or nonlinear functions of the state vector.

The data assimilation results with two minimization algorithms, one being the MLEF and the other being a gradient-based minimization, indicate a clear advantage of the MLEF algorithm. For both the differentiable and non-differentiable observation operators the MLEF was advantageous, showing a robust performance.

Two minimization algorithms based on the MLEF are schematically presented, the generalized nonlinear conjugate-gradient and the generalized quasi-Newton minimization algorithm.   The algorithmic advantage of the generalized minimization methods is that no changes to the original unconstrained minimization algorithms are necessary. Only the calculation of the generalized gradient and of the generalized Hessian

preconditioning is changed to reflect the non-differentiable and nonlinear character of the new method.

The presented MLEF minimization algorithm is directly applicable to ensemble data assimilation methods, and to medium size minimization problems with dimensions of up to $O(10^4)$. For high-dimensional operational applications, for example within the variational data assimilation methods, there are possible extensions of the generalized minimization algorithm using the multi-grid methods and the decomposition into local domains. The particular strategy would depend on the actual minimization problem and the computational environment.

In future work we plan to examine the MLEF performance as a non-differentiable minimization algorithm in more complex applications, such as the assimilation of cloud and microphysics observations, inherently nonlinear and potentially non-differentiable.

Acknowledgments

References

Abramov, R. and A. J. Majda, 2004: Quantifying uncertainty for non-Gaussian ensembles in complex systems. *SIAM J. Sci. Stat. Comp.,* **26**, 411-447.

Akella, S., and Navon, I. M., 2006: A comparative study of the performance of high resolution advection schemes in the context of data assimilation. *Int. J. Numer. Meth. Fluids*, **51**, 719-748.

Anderson, J. L. 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634-642.

Axelsson, O., and Barker, V. A., 1984: *Finite-element solution of boundary value problems: Theory and computation*. Academic Press, 432 pp.

Bishop, B., Etherton, J. and Majumdar, S. J. 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420-436.

Burgers, J. M., 1948: *A Mathematical Model Illustrating the Theory of Turbulence*. Advances in Applied Mechanics, Vol. **1**, Academic Press, New York, 171-199.

Carrio, G.G., Cotton, W. R., Zupanski, D., and M. Zupanski, 2007: Development of an aerosol retrieval method: Description and preliminary tests. *J. Appl. Meteor. Climate* (submitted).

Clarke, F. H., 1983: *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York, 324 pp.

Derber, J.C., 1989: A variational continuous assimilation technique. *Mon.Wea.Rev.*, **117**, 2437-2446.

Evensen, G. 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99** (C5), 10 143-10 162.

Fisher, M., and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast errors in variational data assimilation. *ECMWF Tech. Memo.* 220, 28 pp.

Fletcher, C. A. J., 1991: *Computational Techniques for Fluid Dynamics: Fundamental and General Techniques*. Springer, Berlin, 401 pp.

Fletcher, S. J., and M. Zupanski, 2006a: A data assimilation method for lognormally distributed observation errors. *Quart. J. Roy. Meteor. Soc.*, **132**, 2505-2520.

Fletcher, S. J., and M. Zupanski, 2006b: A hybrid multivariate normal and lognormal distribution for data assimilation. *Atmos. Sci. Let.*, **7**, 43-46.

Gejadze, I., Le Dimet, F. X., and V. Shutyaev, 2007: On error covariances in variational data assimilation. Russian *Journal of Numerical Analysis and Mathematical Modeling*, **22**, 163-175.

Gill, P. E., W. Murray, and M. H. Wright, 1981: *Practical Optimization*. Academic Press, 401 pp.

Greenwald, T. J., R. Hertenstein, and T. Vukicevic, 2002: An all-weather observational operator for radiance data assimilation with mesoscale forecast models. *Mon. Wea. Rev.*, **130**, 1882-1897.

Homescu, C., and I. M. Navon, 2003: Optimal control of flow with discontinuities. *J. Comp. Phys.*, **187**, 660-682.

Houtekamer, P. L. and Mitchell, H. L. 1998: Data assimilation using ensemble

Kalman filter technique. *Mon. Wea. Rev*., **126**, 796-811.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376 pp.

LeDimet, F.X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**, 91-110.

Lemarechal, C., 1977: Bundle methods in nonsmooth optimization. *Proceeding of the IIASA Series*. C. Lemarechal and R. Mifflin, Eds., Pergamon Press, 79-103.

Lemarechal, C., and J. Zowe, 1994: A condensed introduction to bundle methods in nonsmooth optimization. In *Algorithms for Continuous Optimization* (E. Spedicato, ed.), Kluwer Academic Publishers, Dordrecht, 357-382.

Liu, D. C., and J. Nocedal, 1989: On the limited memory BFGS method for large scale optimization. *Math. Programming*, **45**, 503-528.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc*., **112**, 1177-1194.

Luenberger, D. L., 1984: *Linear and Non-linear Programming*. 2nd ed., Addison-Wesley, 491 pp.

Luksan, L., and J. Vlcek, 2001: Algorithm 811: NDA: Algorithms for nondifferentiable Optimization. *ACM Trans. Math. Software*, **27**, 193-213.

Majda, A., K. Haven and R. Abramov, 2004: Quantifying Predictability Through Information Theory: Small Sample Estimation in a Non-Gaussian Framework. *J. Comp. Physics*, **206**, 334–362.

Nesterov, Y., 2005: Smooth minimization of non-smooth functions. *Math. Program*., **103**, 127-152.

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

Nocedal, J., and S. J. Wright, 1999: *Numerical optimization*. Springer-Verlag, New York, 636 pp.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D. J. and Yorke, J. A. 2004: A Local Ensemble Kalman Filter for Atmospheric Data Assimilation. *Tellus*, **56A**, No. 4, 273-277.

Pham, D. T., Verron, J. and Roubaud, M. C. 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Sys*., **16**, 323-340.

Rozier, D., Birol, F., Cosme, E., Brasseur, P., Brankart, J. M., and J. Verron, 2007: A reduced-order Kalman filter for data assimilation in physical oceanography. *SIAM Rev.*, **49**, 449-465.

Tsuyuki, T., 1997: Variational data assimilation in the tropics using precipitation data. Part III: Assimilation of SSM/I precipitation rates. *Mon. Wea. Rev*., **125**, 1447-1464.

Uzunoglu, B., Fletcher, S. J., Zupanski, M., and I. M. Navon, 2007: Adaptive ensemble size inflation and reduction. *Quart. J. Roy. Meteor. Soc*., **133**, 1281-1294.

van Leeuwen, P.J., 2003: A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev*., **131**, 2071–2084.

Veerse, F., 1999: Variable-storage quasi-Newton operators as inverse forecast/analysis

error covariance matrices in variational data assimilation. *INRIA Research Report 3685*, p. 28.

Verlinde, J., and W. R. Cotton, 1993: Fitting microphysical observations of non-steady convective clouds to a numerical model: An application of the adjoint technique of data assimilation to a kinematic model. *Mon. Wea. Rev.*, **121**, 2276–2793.

Wang, X., Bishop, C. H., and S. J. Julier, 2004: Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590-1605.

Wei, M., Toth, Z., Wobus, R., Zhu, Y., Bishop, C. H., and X. Wang, 2006: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP, *Tellus*, **58A**, 28-44.

Whitaker, J. S., and Hamill, T. M. 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**,1913-1924.

Xu, Q., and J. Gao. 1999: Generalized adjoint for physical processes with parameterized discontinuities. Part VI: Minimization problems in multidimensional space. *J. Atmos. Sci.*, **56**, 994–1002.

Zhang, S., X. Zou, J. Ahlquist, I. M. Navon, and J. G. Sela, 2000: Use of differentiable and nondifferentiable optimization algorithms for variational data assimilation with discontinuous cost functions. *Mon. Wea. Rev.*, **128**, 4031-4044.

Zupanski, D., 1993: The effects of discontinuities in the Betts-Miller cumulus convection scheme on four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Tellus*, **45A**, 511-524.

Zupanski, D. and Zupanski, M. 2006: Model error estimation employing ensemble data assimilation approach. *Mon. Wea. Rev.*, **134**, 1337-1354.

Zupanski, D. Denning, A. S., Uliasz, M., Zupanski, M., Schuh, A. E., Rayner, P. J., Peters, W., and K. D. Corbin, 2007a: Carbon flux bias estimation employing Maximum Likelihood Ensemble Filter (MLEF). *J. Geophys. Res.*, **112**, D17107, doi:10.1029/2006JD008371.

Zupanski, D., Hou, A. Y., Zhang, S.Q., Zupanski, M., Kummerow, and S. H. Cheung, 2007b: Applications of information theory in ensemble data assimilation. *Quart. J. Roy. Meteor. Soc.*, (in press).

Zupanski, M. 2005: Maximum Likelihood Ensemble Filter: Theoretical Aspects. *Mon. Wea. Rev.*, **133**, 1710–1726.

Zupanski, M., Fletcher, S. J., Navon, I. M., Uzunoglu, B., Heikes, R. P., Randall, D. A., Ringler, T. D., and D. Daescu, 2006: Initiation of ensemble data assimilation. *Tellus*, **58A**, 159-170.
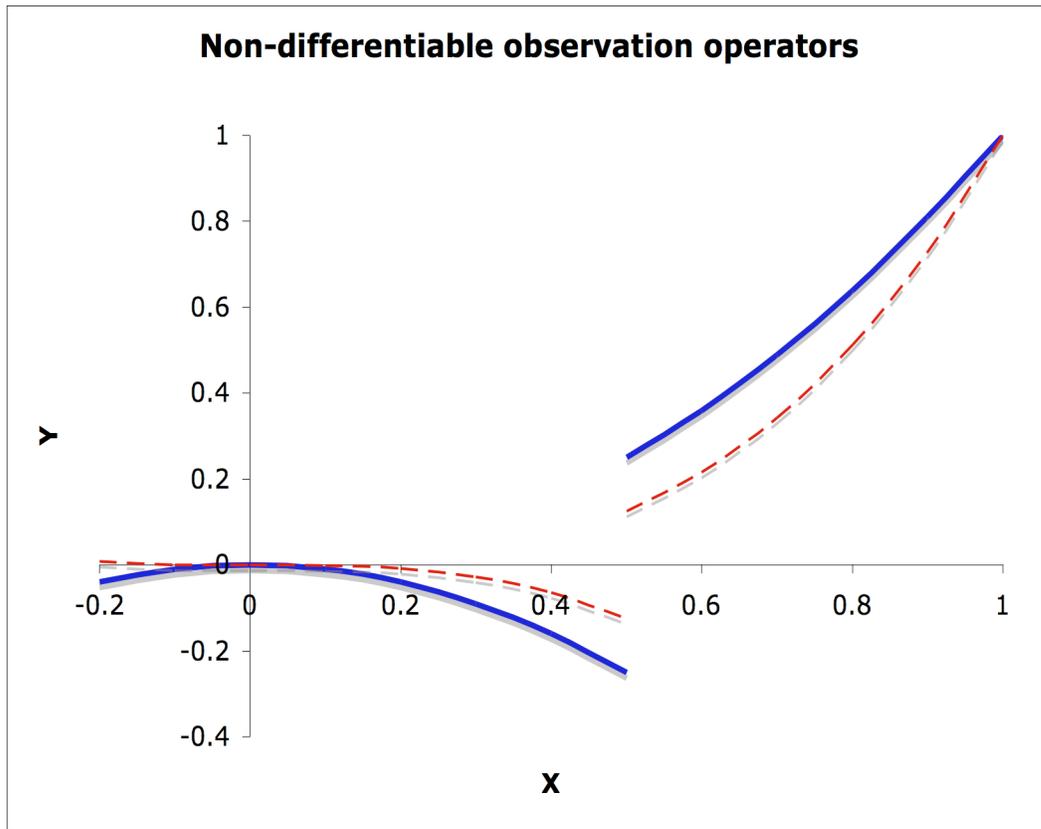
Figure 1. Non-differentiable observation operators: (*a*) quadratic (full line) and (*b*) cubic (dashed line). The discontinuous point at $x=0.5$ represent a discontinuity in the function and its first and second derivatives.
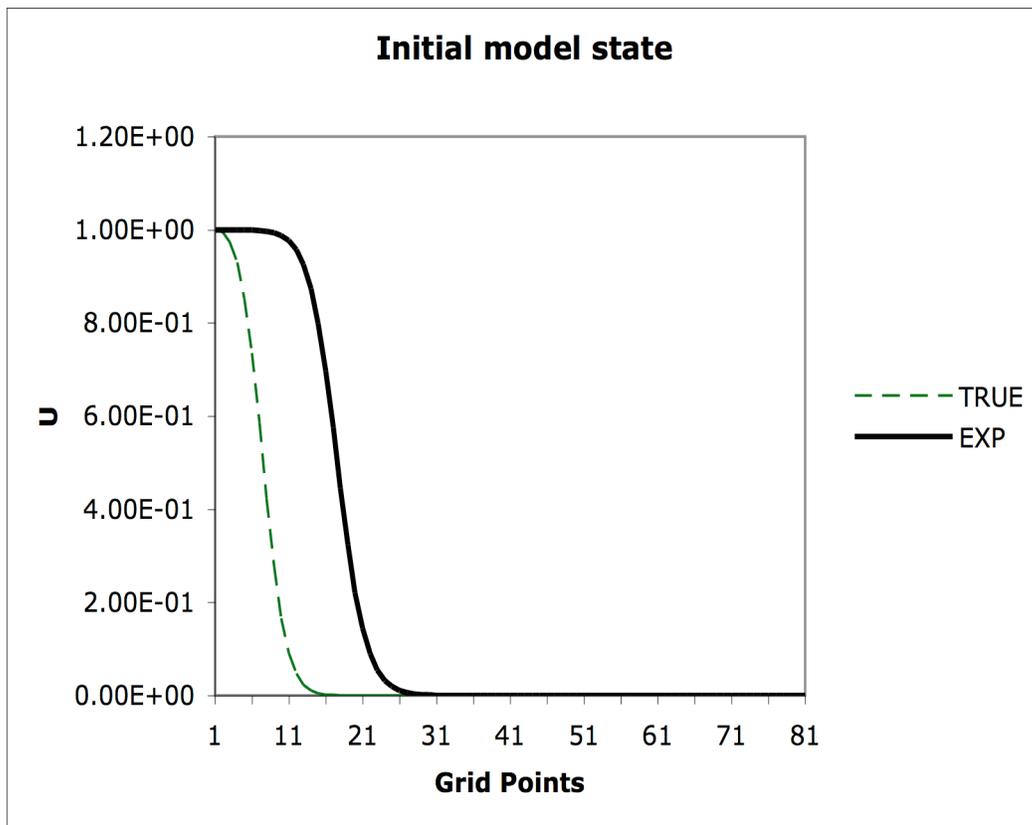
Figure 2. The 'truth' (TRUE) and the 'experimental' (EXP) initial conditions before data assimilation. The true state (dashed line) was used to initiate the true forecast from which the observations are created, while the initial conditions in the experiment are defined as a 20-point shift of the truth (solid line), obtained as a lagged forecast.

(*a*)
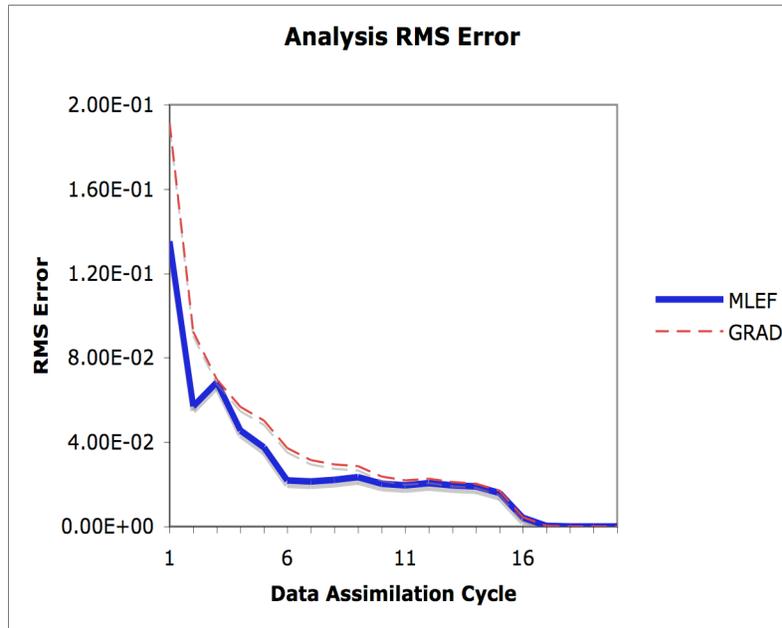
(*b*)
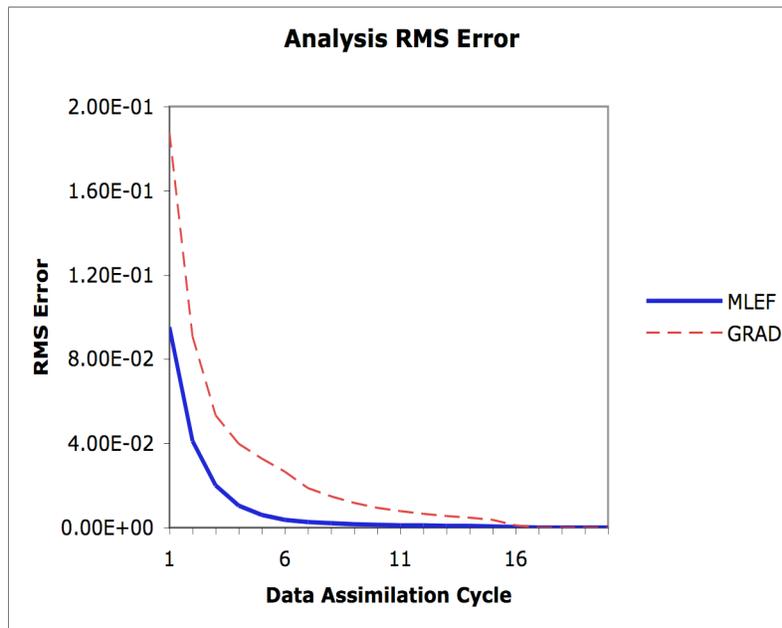
Figure 3. Analysis root-mean-squared (RMS) error in the non-differentiable observation operator examples: (a) quadratic, and (b) cubic. The MLEF results are represented by solid line, while the gradient-based method results (GRAD) are represented by dashed line.
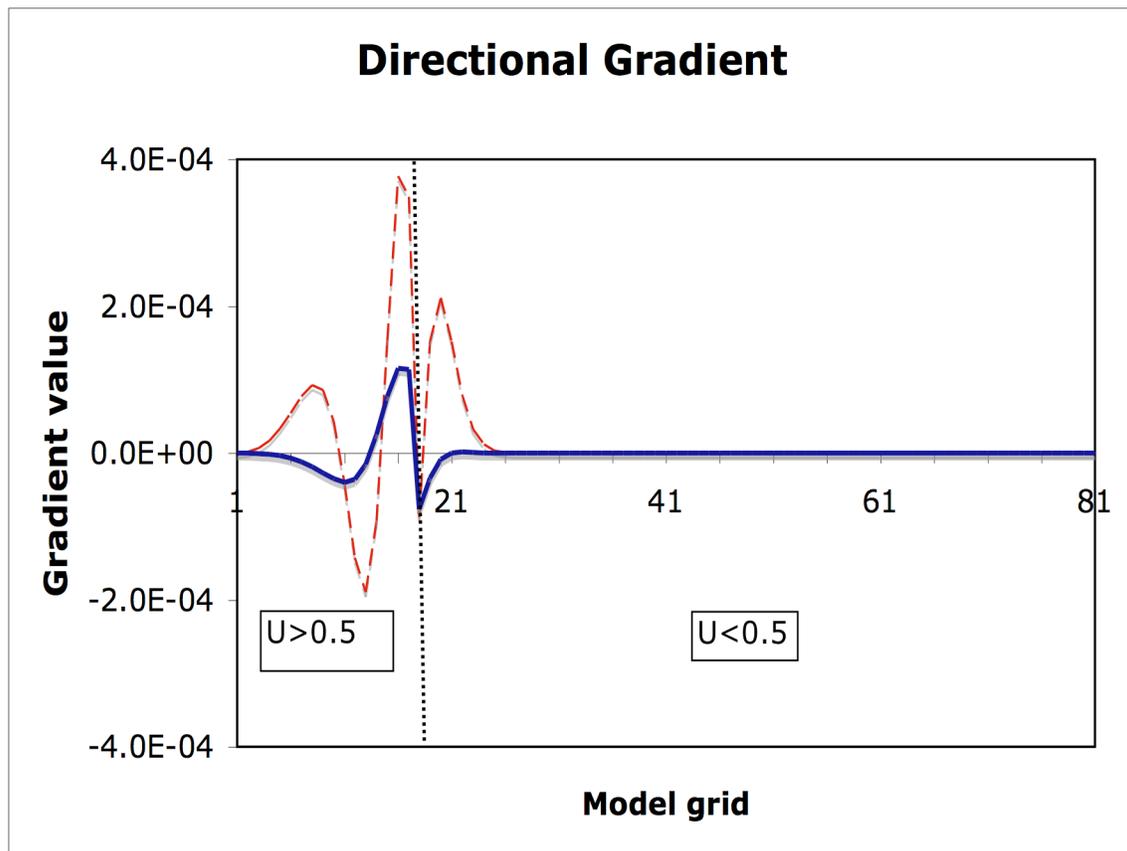
Figure 4. Directional gradient in the GRAD (dashed line) and the MLEF (solid line) experiments. The discontinuous point at model value U=0.5 is denoted by a vertical dotted line. The encapsulated text boxes define the values of the model variable U.

(a)

(b)

(c)

(d)
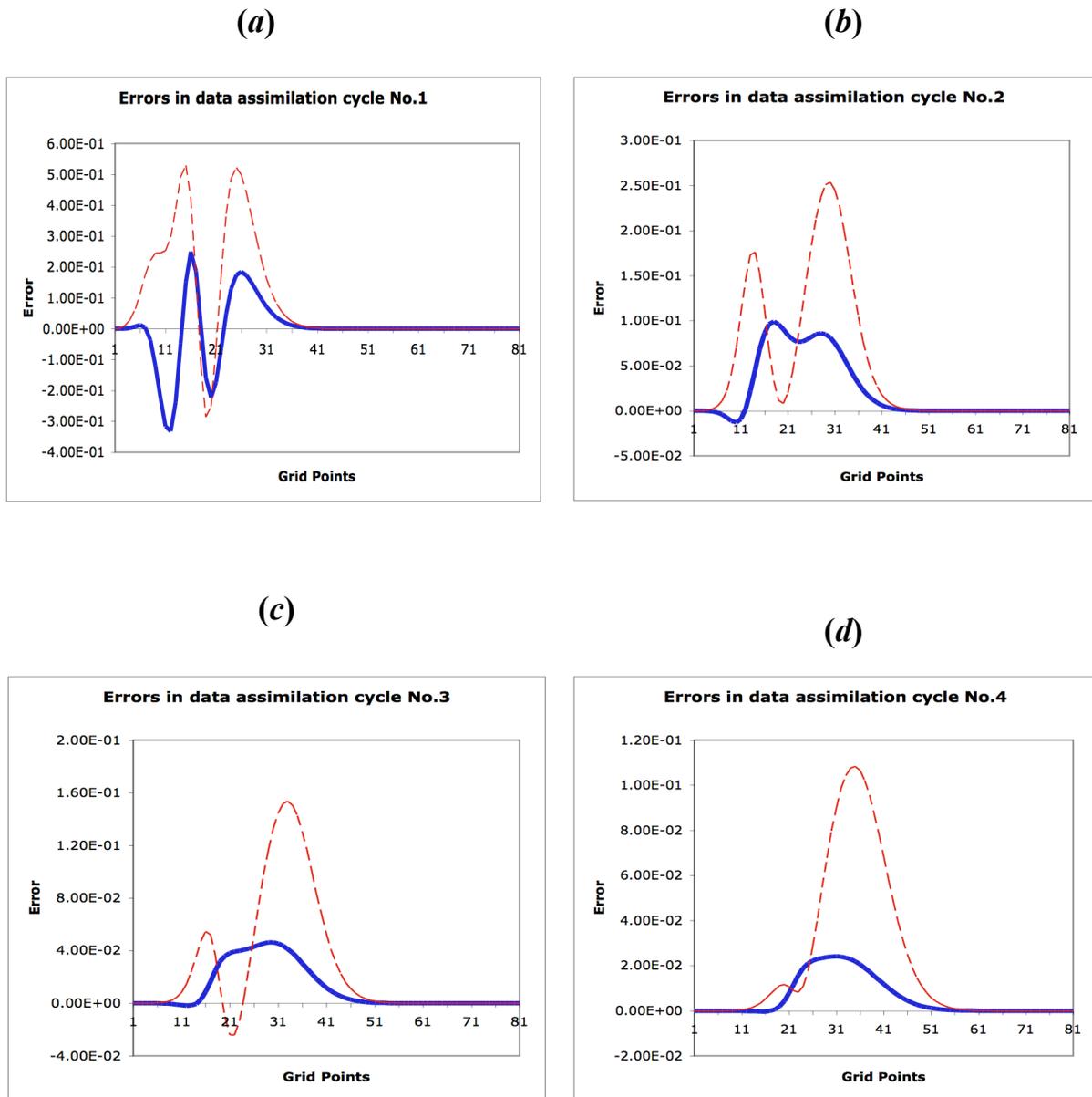
Figure 5. The state vector analysis errors $x^a - x^t$ in first four data assimilation cycles, for the MLEF (solid line) and the GRAD (dashed line) experiments: (a) after cycle No.1, (b) after cycle No.2, (c) after cycle No.3, and (d) after cycle No.4.

(a)

(b)
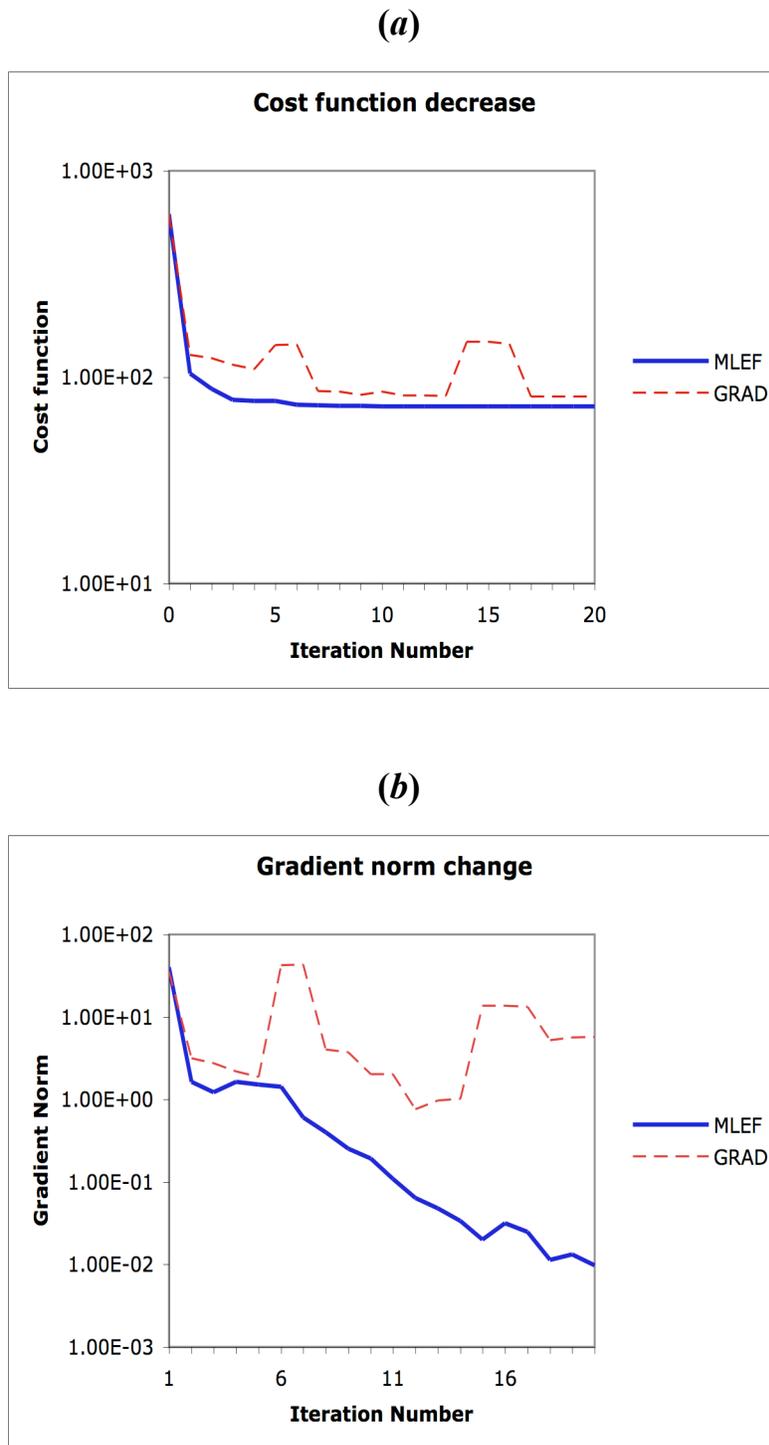
Figure 6. Minimization in the quadratic non-differentiable observation operator example: (a) cost-function, and (b) gradient norm. The MLEF results are represented by solid line, and the gradient-based method results (GRAD) are represented by dashed line.

(*a*)
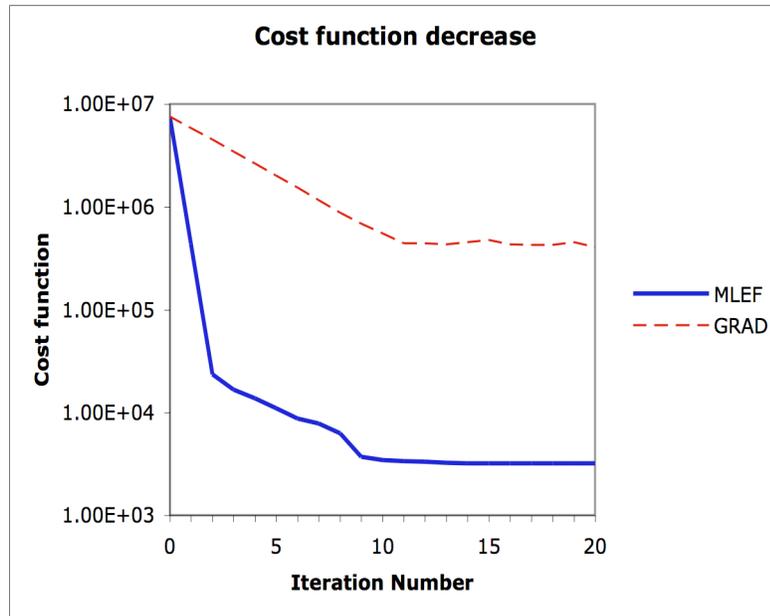
(*b*)
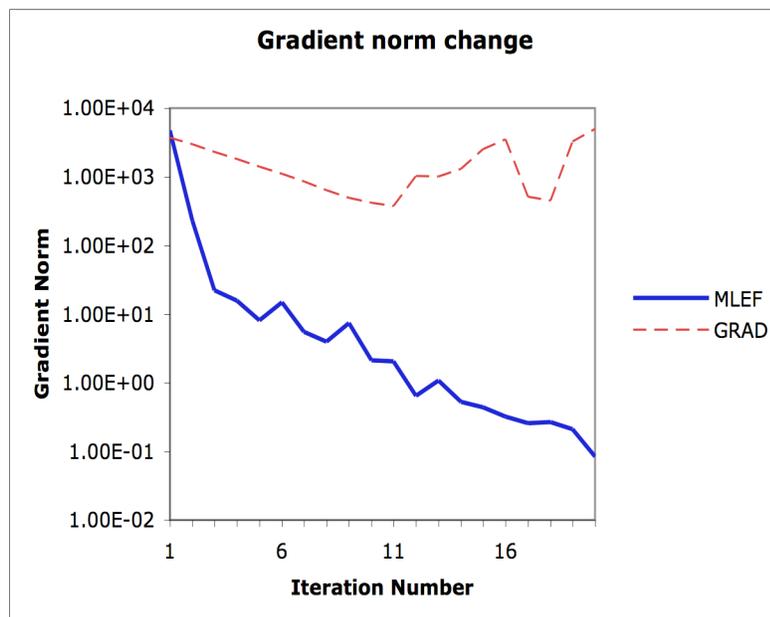
Figure 7. Same as in Fig.6, except in the cubic non-differentiable observation operator example.

(a)
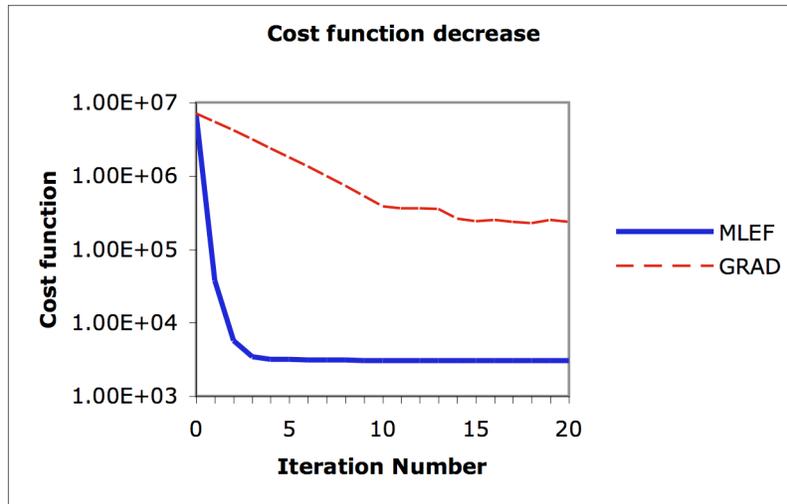
(b)
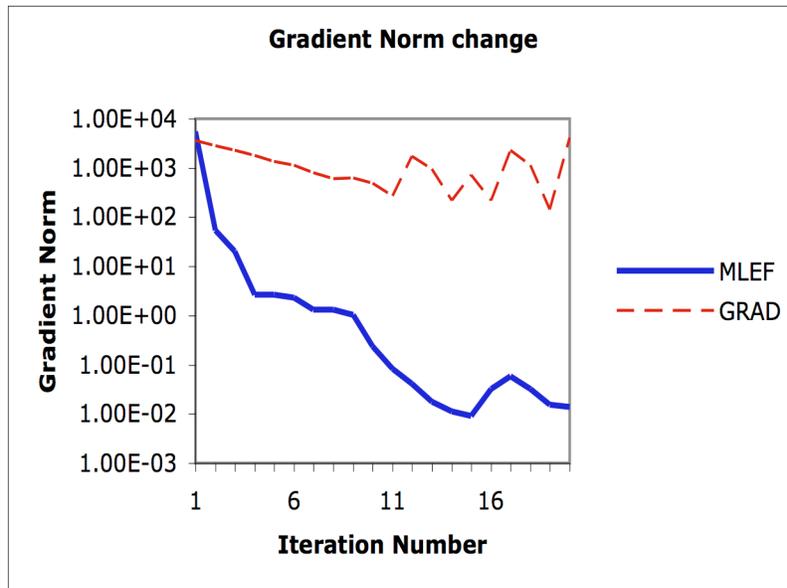
Figure 8. Same as in Figure 7, except for differentiable observation operator.