Q. J. R. Meteorol. Soc. (2001), **127**, pp. 661–683

Optimality of variational data assimilation and its relationship with the Kalman filter and smoother

By ZHIJIN LI and I. M. NAVON*
Florida State University, USA

(Received 23 June 1999; revised 25 August 2000)

SUMMARY

The known properties of equivalence between four-dimensional variational (4D-Var) data assimilation and the Kalman filter as well as the fixed-interval Kalman smoother point to particular optimal properties of 4D-Var. In the linear context, the 4D-Var solution is optimal, not only with respect to the model trajectory segment over the assimilation time interval, but also with respect to any model state at a single observation time level; in the batch processing (cycling 4D-Var) method, the information in 4D-Var is fully transferred from one batch to the next by the background term; 4D-Var allows the processing of observations in subsets, while the final solution is optimal as all observations are processed simultaneously. These properties hold even for models that are imperfect, as well as not invertible. Various properties of equivalence of 4D-Var to the Kalman filter and smoother result from these optimality properties of 4D-Var. Further, we show that the fixed-lag Kalman smoother may also be constructed in an optimal way using a multiple batch-processing 4D-Var approach. While error covariances are crucial for the equivalence, practical techniques for evaluating error covariances in the framework of cycling 4D-Var are discussed.

KEYWORDS: Data assimilation Kalman filter Kalman smoother Optimality

1. INTRODUCTION

Methodologies of meteorological data assimilation comprise two families: variational data assimilation and sequential data assimilation (Daley 1991; Ghil and Malanotte-Rizzoli 1991; Cohn 1997; Courtier 1997; Ghil 1997). Both variational and sequential data assimilation fit into the framework of estimation theory, though variational methods are closer to aspects of control theory (Lorenc 1986; Ghil and Malanotte-Rizzoli 1991; Cohn 1997; and references therein). The variational data assimilation methods include both three-dimensional and four-dimensional variational data assimilation (3D-Var and 4D-Var respectively), while the sequential data assimilation features the Kalman filter and smoother.

Variational data assimilation has undergone rapid development since the work by Le Dimet and Talagrand (1986), Lewis and Derber (1985), Talagrand and Courtier (1987), and Courtier and Talagrand (1987). 3D-Var was put in operational use in place of the optimal interpolation (OI) method in 1991 at the National Centers for Environment Prediction (NCEP) (Derber *et al.* 1991; Parrish and Derber 1992). Since then, most meteorological centres in the world have implemented 3D-Var, including the European Centre for Medium-Range Weather Forecasts (ECMWF) in 1996 (Courtier *et al.* 1998), the Canadian Meteorological Center in 1997 (Gauthier *et al.* 1998), the Data Assimilation Office (Cohn *et al.* 1998), the Met Office of the UK (UKMO) (Ingleby *et al.* 1999), Météo-France (Thépaut *et al.* 1998), and the High-Resolution Limited-Area Model (HIRLAM) group (Berre *et al.* 1999). At the same time, 4D-Var is intensively investigated and developed in conjunction with operational models (e.g. Thépaut and Courtier 1991; Navon *et al.* 1992; Chao and Chang 1992; Zupanski 1993; Courtier *et al.* 1994; Zou and Kuo 1996; Zupanski and Mesinger 1995; to cite just a few). The most encouraging achievement occurred recently at ECMWF, where 4D-Var was implemented for operational use at the end of 1997. 4D-Var was implemented using numerical weather prediction (NWP) models even with the most sophisticated physical

* Corresponding author: School of Computational Science and Information Technology, Florida State University, Tallahassee, FL32306-4120, USA.

processes (e.g. Zou and Kuo 1996; Zupanski 1997; Mahfouf and Rabier 2000; Rabier *et al.* 1998, 2000).

Parallel with variational methods, sequential methods have been extensively investigated since the beginning of the 1980's (Ghil *et al.* 1981). Most meteorological applications of the Kalman filter are limited to idealized one- and two-dimensional models (e.g. Cohn and Parrish 1991; Dee 1991; Daley 1992; Evensen 1994; Todling and Cohn 1994). Cohn *et al.* (1994) introduced a Kalman smoother to meteorology, named retrospective data assimilation, aiming to produce a long time-consistent sequence of analyses. This retrospective method has also been studied using idealized one- and two-dimensional models (Cohn *et al.* 1994; Todling *et al.* 1998).

Variational methods seem to dominate the scene of meteorological data assimilation. In estimation theory, variational methods are called statistical methods or the least-squares approach (Jazwinski 1970, p. 51). The least-squares approach does not seem to contribute much to estimation theory. The current popularity of variational methods can primarily be attributed to two facts, namely, the large size of atmospheric models and the availability of efficient minimization algorithms associated with adjoint techniques (Le Dimet and Talagrand 1986; Talagrand and Courtier 1987). Contrary to variational methods, the large size of meteorological models has prevented the Kalman filter and smoother from being implemented in their full form. Some significant advances have been made in developing sub-optimal Kalman filters and smoothers (e.g. Todling *et al.* 1998), but enormous efforts are still required to operationally implement these sub-optimal systems in meteorology.

Although sequential methods and variational methods are implemented separately, they are intimately connected to each other. Some connections between them have long been known. For a perfect, linear model and linear observation operators, both 4D-Var and the Kalman filter (Kalman 1960) yield the same values for the model variables at the end of the 4D-Var assimilation period (e.g. Lorenc 1986; Ghil 1989). 4D-Var and the fixed-interval Kalman smoother are equivalent even for models with errors (Bryson and Ho 1975; Bennett and Budgell 1989; Ménard and Daley 1996).

The equivalence between sequential and variational methods provides a possibility of combining these two methods, which may result in more powerful algorithms. A few efforts in this direction have already been undertaken. A major effort is related to cycling 3D-Var and 4D-Var (Cohn 1993; Courtier *et al.* 1994). Cycling 3D-Var is algorithmically equivalent to the Kalman filter, while cycling 4D-Var operates in a way similar to the fixed-lag Kalman smoother. Recently, it has been emphatically suggested that simplified Kalman filters can be used for evaluating error covariance matrices required by 3D-Var and 4D-Var (Ménard and Daley 1996; Ehrendorfer and Bouttier 1998). These methods have shown prospects for practical use (Fisher 1998). We know that the solutions of the Kalman filter/smoother are optimal in the sense of the maximum likelihood or minimum error variance. It is easy to show that cycling 3D-Var is exactly the Kalman filter in the framework of a perfect model (e.g. Lorenc 1986; Rabier *et al.* 1993). However, the optimality of cycling 4D-Var has not previously been shown, especially when model errors are taken into consideration.

Thus, in this work, we intend to present a consistent and detailed analysis of the optimality of 4D-Var solutions. Then, the issues of the optimality of cycling 4D-Var and the equivalences between the sequential and variational methods are further investigated.

It is known that a 4D-Var solution is optimal in the sense of *joint* maximum likelihood (Bayesian) estimation, i.e. maximizing the joint conditional density function. It has also been recognized that the 4D-Var solution at the end of the time window is optimal in the sense of maximizing the marginal conditional density function, and thus

is equivalent to the Kalman filter solution (Jazwinski 1970; Lorenc 1986; Ghil 1989). Here we will extend these results by proving that the 4D-Var solution is optimal for any single observation time level in the sense of maximizing the marginal conditional density function, a property called *the consistent optimality property.*

We will further show that the information in 4D-Var can be fully transferred from one batch to the next by a background term when batch-processing (cycling 4D-Var) methods are used, called *the transferable optimality property.* Also, 4D-Var allows the processing of observations in subsets, while the final solution is optimal as all observations are processed simultaneously, and such a property is called *the additive optimality property.* Actually, these two optimal properties have been used in the past but only as *ad hoc* assumptions. The consistent and transferable properties render cycling 4D-Var to be as optimal as a fixed-lag Kalman smoother when the exact error covariance is available. The additive property implies an exact 4D-Var solution when computed by quasi-continuous variational data assimilation (Jarvinnen *et al.* 1996). It is noteworthy that these three properties are derived for perfect models as well as models with errors, but for linear systems with Gaussian distributions.

These optimality properties of 4D-Var can be used to demonstrate various connections between 4D-Var and the Kalman filter/smoother. As we have mentioned, the relationship between 4D-Var and the Kalman filter has long been known (Jazwinski 1970; Lorenc 1986; Ghil 1989), and also the equivalence between 4D-Var and the fixed-interval Kalman smoother (e.g. Bryson and Ho 1975; Bennett and Budgell 1989; Ménard and Daley 1996). We will show here that the batch (cycling) 4D-Var is equivalent to the fixed-lag Kalman smoother from the 4D-Var consistent and transferable optimality properties. We learned that Zhu *et al.* (1999) have comprehensively examined the relationship between 4D-Var and smoother algorithms when we submitted this paper. Zhu *et al.* (1999) have also shown the equivalence of 4D-Var and the fixed-lag Kalman smoother by using a sweep method (Bryson and Ho 1975; Ménard and Daley 1996), but the constructive procedure of the fixed-lag Kalman smoother is different from the one presented here.

The derivation of the above results has led to the presentation of some new formulations. These formulations are useful for further theoretical analysis and for designing cycling 4D-Var/3D-Var. Actually, our derivation is 4D-Var algorithm oriented, which is especially emphasized when model errors are considered.

The outline of this paper is as follows. In section 2, we briefly summarize the mathematical assumptions used in this study. Section 3 presents some properties of 4D-Var. The optimality of 4D-Var is detailed in section 4. Section 5 examines the equivalence between the variational and sequential methods, and a 4D-Var smoother is proposed for producing a time-consistent long sequence of assimilated analyses. Possible strategies for evaluating error covariances are analysed in section 6. Finally, section 7 discusses and summarizes the results derived in this study.

2. Basic assumptions and mathematical formulations

In what follows, we will restrict our analysis to linear dynamics and observational operators. Consider the discrete stochastic dynamical system described by the stochastic vector difference equation,

$$\mathbf{x}_{k+1} = \mathbf{L}(k+1, k)\mathbf{x}_k + \Gamma(k)\mathbf{w}_{k+1}, \quad k = 0, 1, \ldots, \tag{1}$$

where the state at t_k is $\mathbf{x}_k$, an n-vector, $\mathbf{L}(k+1, k)$ is an $n \times n$ matrix, $\Gamma(k)$ is $n \times r$, and $\mathbf{w}_k$ $\{k = 1, \ldots\}$ is an r-vector, white Gaussian sequence, i.e.

$$\mathbf{w}_k \sim \mathrm{N}(0, \mathbf{Q}_k), \tag{2}$$

where N stands for the Gaussian (normal) distribution, and $\mathbf{Q}_k$ is the covariance of $\mathbf{w}_k$. The distribution of the initial condition $\mathbf{x}_0$ is assumed given,

$$\mathbf{x}_0 \sim \mathrm{N}(\mathbf{x}_0^{\mathrm{b}}, \mathbf{B}_0), \tag{3}$$

where $\mathbf{x}_0^{\mathrm{b}}$ is the background state, $\mathbf{B}_0$ is the analysis error covariance matrix at time t_0, and $\mathbf{x}_0$ is independent of $\{\mathbf{w}_k\}$. We can see that the random sequence generated by Eq. (1) is Markov, since $\mathbf{x}_{k+1}$ is determined by $\mathbf{x}_k$ and $\mathbf{w}_k$, independently of $\{\mathbf{x}_{k-1}, \ldots, \mathbf{x}_0\}$ (Jazwinski 1971, p. 86). Let discrete, noisy, m-vector observations (measurements) $\mathbf{y}_k$ be given by

$$\mathbf{y}_k = \mathbf{h}_k \mathbf{x}_k + \mathbf{v}_k \tag{4}$$

where $\mathbf{h}_k$ is an $m \times n$ matrix and $\{\mathbf{v}_k, k = 1, \ldots\}$ is an m-vector, white Gaussian sequence, $\mathbf{v}_k \sim \mathrm{N}(0, \mathbf{R}_k)$, $\mathbf{R}_k > 0$. $\{\mathbf{w}_k\}$ and $\{\mathbf{v}_k\}$ are assumed independent, and $\{\mathbf{w}_k\}$ is independent of $\mathbf{x}_0$. We note that the joint $\{\mathbf{x}_k, \mathbf{y}_k\}$ process is Markov.

Let $\mathbf{Y}_l$ be the sequence of observations

$$\mathbf{Y}_l = \{\mathbf{y}_1, \ldots, \mathbf{y}_l\}. \tag{5}$$

Given a realization of the sequence of observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$, that is, given $\mathbf{Y}_l$, the estimation problem consists of computing an estimate $\mathbf{x}_k$ based on $\mathbf{Y}_l$. If $k < l$, the problem is called the smoothing problem; if $k = l$, it is called the filtering problem; and if $k > l$, it is called the prediction problem.

Let $\hat{\mathbf{x}}_k$ be an estimate of $\mathbf{x}_k$ given $\mathbf{Y}_l$. The goal of data assimilation is to obtain the optimal $\hat{\mathbf{x}}_k$. It is necessary to define criteria measuring the optimality. We have the error of estimation

$$\delta \mathbf{x}_k \equiv \mathbf{x}_k^{\mathrm{t}} - \hat{\mathbf{x}}_k, \tag{6}$$

where $\mathbf{x}_k^{\mathrm{t}}$ is the true state. The criterion should be that the optimal estimate has the smallest error.

Estimation theory shows that the minimum mean square error or minimum variance estimate is the conditional mean, that is, the mean conditioned by observations (Jazwinski 1970; Cohn 1997). It is noteworthy that this result is general and independent of the nature of the conditional probability density $p(\mathbf{x}_k|\mathbf{Y}_l)$ (Cohn 1997). Thus, the conditional mean is a reasonable optimal estimate.

The alternative is the maximum likelihood (Bayesian) estimate obtained by maximizing the conditional probability density function $p(\mathbf{x}_k|\mathbf{Y}_l)$. When $p(\mathbf{x}_k|\mathbf{Y}_l)$ is Gaussian, the maximum likelihood (Bayesian) estimate is identical to the minimum variance estimate. In the following, we shall use terminologies associated with the maximum likelihood (Bayesian) estimate to analyse both the variational and sequential methods (e.g. Jazwinski 1970).

3. Properties of four-dimensional variational data assimilation

(a) *Perfect model*

(i) *Consistent properties of the time interval.* The standard 4D-Var minimizes the cost (objective) function J that measures the weighted sum of squares of distances to the

background state $\mathbf{x}^{\mathrm{b}}$ and to the observations $\mathbf{y}$ distributed over a time interval (t_0, t_l) (Lorenc 1986),

$$\mathrm{J}^{\mathrm{p}} = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}) + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k) \tag{7}$$

where the p superscript refers to the perfect model. Observations start from time t_1. This is only for the sake of convenience when we describe cycling 4D-Var.

In the standard 4D-Var, the minimization of the cost function Eq. (7) is carried out with respect to the initial state $\mathbf{x}_0$. As such, the cost function Eq. (7) becomes

$$\begin{aligned}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0) = {} & \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}) \\ & + \frac{1}{2}\sum_{k=1}^{l}\{\mathbf{h}_k\mathbf{L}(k, 0)\mathbf{x}_0 - \mathbf{y}_k\}^{\mathrm{T}}\mathbf{R}_k^{-1}\{\mathbf{h}_k\mathbf{L}(k, 0)\mathbf{x}_0 - \mathbf{y}_k\}.\end{aligned} \tag{8}$$

Here, we have used the notation

$$\mathbf{L}(k, i) = \mathbf{L}(k, k-1)\ldots\mathbf{L}(i+1, i), \quad \text{for } i \le k. \tag{9}$$

For the cost function with respect to $\mathbf{x}_0$, the analytical expression of the optimal solution has been presented by, for example, Lorenc (1986) and Thépaut and Courtier (1991). Here, we still present the basic expressions for the sake of the paper being self-contained.

Differentiating Eq. (8), we obtain the gradient of $\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0)$ with respect to $\mathbf{x}_0$

$$\nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0) = \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}) + \sum_{k=1}^{l}\mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\{\mathbf{h}_k\mathbf{L}(k, 0)\mathbf{x}_0 - \mathbf{y}_k\}. \tag{10}$$

A minimum solution satisfies $\nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0) = 0$. From Eq. (10), we then obtain the optimal solution

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0^{\mathrm{b}} - \mathbf{H}_{0,l}^{-1}\nabla_{\mathbf{x}_0}\mathrm{J}(\mathbf{x}_0^{\mathrm{b}}), \tag{11}$$

where

$$\mathbf{H}_{0,l} = \mathbf{B}_0^{-1} + \sum_{k=1}^{l}\mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k, 0) \tag{12}$$

is the second derivative of the cost function Eq. (8) at $\mathbf{x}_0$, called the Hessian matrix, and

$$\nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0^{\mathrm{b}}) = \sum_{k=1}^{l}\mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k^{\mathrm{b}} - \mathbf{y}_k), \tag{13}$$

is the gradient of the cost function Eq. (8) at $\mathbf{x}_0^{\mathrm{b}}$. Here

$$\mathbf{y}_k^{\mathrm{b}} = \mathbf{h}_k\mathbf{x}_k^{\mathrm{b}} = \mathbf{h}_k\mathbf{L}(k, 0)\mathbf{x}_0^{\mathrm{b}}. \tag{14}$$

The Hessian matrix is symmetric and positive semi definite.

The error covariance matrix is (Rabier and Courtier 1992)

$$\mathbf{P}_0^{\mathrm{a}} = \langle(\hat{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{t}})(\hat{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{t}})^{\mathrm{T}}\rangle = \mathbf{H}_{0,l}^{-1}, \tag{15}$$

where $\mathbf{x}_0^t$ is the assumed true value of $\mathbf{x}_0$. By using the Sherman–Morrison–Woodbury formula ((7B.5) in Jazwinski (1970)), we can easily get the expression given by Lorenc (1986) and Thépaut and Courtier (1991) in terms of the Kalman gain (Kalman 1960; Cohn 1997). It is of practical importance that the analysis error covariance matrix is the inverse of the Hessian matrix (see section 6).

We now turn to our concern about the optimality of each state on the trajectory determined by $\hat{\mathbf{x}}_0$ over the entire time window. That is, we want to prove that $\mathbf{L}(k, 0)\hat{\mathbf{x}}_0$ minimizes Eq. (7). To prove this we only need to demonstrate that $\hat{\mathbf{x}}_k = \mathbf{L}(k, 0)\hat{\mathbf{x}}_0$ $(k = 1, \ldots, l)$ minimizes the following cost function:

$$\begin{aligned} \mathrm{J}^{\mathrm{p}}(\mathbf{x}_k) = & \frac{1}{2}(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}})^{\mathrm{T}}\mathbf{L}^{-\mathrm{T}}(k, 0)\mathbf{B}_0^{-1}\mathbf{L}^{-1}(k, 0)(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}}) \\ & + \frac{1}{2}\sum_{i=1}^{k}\{\mathbf{h}_i\mathbf{L}^{-1}(k, i)\mathbf{x}_k - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{R}_i^{-1}\{\mathbf{h}_i\mathbf{L}^{-1}(k, i)\mathbf{x}_k - \mathbf{y}_i\} \\ & + \frac{1}{2}\sum_{i=k+1}^{l}\{\mathbf{h}_i\mathbf{L}(i, k)\mathbf{x}_k - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{R}_i^{-1}\{h_i\mathbf{L}(i, k)\mathbf{x}_k - \mathbf{y}_i\}. \end{aligned} \tag{16}$$

Here $\mathbf{L}^{-\mathrm{T}}$ stands for $(\mathbf{L}^{\mathrm{T}})^{-1}$.

This cost function involves the inverse of $\mathbf{L}(k, 0)$. An atmospheric model may not necessarily be invertible. Thus, we introduce the generalized or Moore–Penrose inverse

$$\tilde{\mathbf{L}}^{-1}(k, 0) = \mathbf{V}(k, k-1)\tilde{\Sigma}^{-1}(k, 0)\mathbf{U}^{\mathrm{T}}(k, 0). \tag{17}$$

Actually, the generalized inverse is defined by the singular value decomposition

$$\mathbf{L}(k, 0) = \mathbf{U}(k, 0)\Sigma(k, 0)\mathbf{V}^{\mathrm{T}}(k, 0), \tag{18}$$

where $\Sigma(k, 0)$ is a diagonal matrix, and both $\mathbf{U}(k, 0)$ and $\mathbf{V}(k, 0)$ are orthonormal matrices (see appendix for the definition of $\Sigma(k, 0)$, $\mathbf{U}(k, 0)$ and $\mathbf{V}(k, 0)$). Let $\Sigma(k, 0) = \mathrm{diag}(\sigma_1, \ldots, \sigma_i, \ldots, 0, \ldots, 0)$. Then

$$\tilde{\Sigma}^{-1}(k, 0) = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_i^{-1}, \ldots, 0, \ldots, 0). \tag{19}$$

The cost function then becomes

$$\begin{aligned} \mathrm{J}^{\mathrm{p}}(\mathbf{x}_k) = & \frac{1}{2}(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}})^{\mathrm{T}}\tilde{\mathbf{L}}^{-\mathrm{T}}(k, 0)\mathbf{B}_0^{-1}\tilde{\mathbf{L}}^{-1}(k, 0)(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}}) \\ & + \frac{1}{2}\sum_{i=1}^{k}\{\mathbf{h}_i\tilde{\mathbf{L}}^{-1}(k, i)\mathbf{x}_k - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{R}_i^{-1}\{\mathbf{h}_i\tilde{\mathbf{L}}^{-1}(k, i)\mathbf{x}_k - \mathbf{y}_i\} \\ & + \frac{1}{2}\sum_{i=k+1}^{l}\{\mathbf{h}_i\mathbf{L}(i, k)\mathbf{x}_k - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{R}_i^{-1}\{h_i\mathbf{L}(i, k)\mathbf{x}_k - \mathbf{y}_i\}. \end{aligned} \tag{20}$$

We can deduce that $\mathrm{J}^{\mathrm{p}}(\mathbf{x}_k)$ is simply $\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0)$ by using $\mathbf{x}_0 = \tilde{\mathbf{L}}^{-\mathrm{T}}(k, 0)\mathbf{x}_k$ and $\tilde{\mathbf{L}}^{-1}(k, 0)\mathbf{L}(k, 0) = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Using the chain rule, we obtain the gradient with respect to $\mathbf{x}_k$

$$\nabla_{\mathbf{x}_k}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_k) = \tilde{\mathbf{L}}^{-1}(k, 0)\nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0). \tag{21}$$

From the optimality condition $\nabla_{\mathbf{x}_k}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_k) = 0$, some algebraic manipulations yield

$$\hat{\mathbf{x}}_k = \mathbf{x}_k^{\mathrm{b}} - \mathbf{H}_{k,l}^{-1}\nabla_{\mathbf{x}_k}\mathrm{J}(\mathbf{x}_k^{\mathrm{b}}), \tag{22}$$

where

$$\mathbf{H}_{k,l} = \tilde{\mathbf{L}}^{-\mathrm{T}}(k,0)\mathbf{B}_0^{-1}\tilde{\mathbf{L}}^{-1}(k,0) + \sum_{i=1}^{k} \tilde{\mathbf{L}}^{-\mathrm{T}}(k,i)\mathbf{h}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{h}_i\tilde{\mathbf{L}}^{-1}(k,i)$$

$$+ \sum_{i=k+1}^{l} \mathbf{L}^{\mathrm{T}}(i,k)\mathbf{h}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{h}_i\mathbf{L}(i,k) \tag{23}$$

is the Hessian matrix of the cost function Eq. (16) at $\mathbf{x}_k$, and

$$\nabla_{\mathbf{x}_k}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_k^{\mathrm{b}}) = \tilde{\mathbf{L}}^{-1}(k,0)\nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0^{\mathrm{b}}) \tag{24}$$

is the gradient of the cost function Eq. (16) at $\mathbf{x}_k^{\mathrm{b}}$.

A procedure similar to that used to obtain Eq. (15) (Rabier and Courtier 1992) yields

$$\mathbf{P}_k^{\mathrm{a}} = \langle(\hat{\mathbf{x}}_k - \mathbf{x}_k^{\mathrm{t}})(\hat{\mathbf{x}}_k - \mathbf{x}_k^{\mathrm{t}})^{\mathrm{T}}\rangle = \mathbf{H}_{k,l}^{-1}, \tag{25}$$

where $\mathbf{x}_k^{\mathrm{t}}$ is the true value of $\mathbf{x}_k$.

We can now associate Eqs. (22), (23), (24), and (25) with the counterparts of the optimal solution $\mathbf{x}_0$.

$$\hat{\mathbf{x}}_k = \mathbf{L}(k,0)\hat{\mathbf{x}}_0 \tag{26}$$

$$\mathbf{L}^{\mathrm{T}}(k,0)\mathbf{H}_{k,l}\mathbf{L}(k,0) = \mathbf{H}_{0,l} \tag{27}$$

$$\mathbf{L}^{\mathrm{T}}(k,0)\nabla_{\mathbf{x}_k}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_k^{\mathrm{b}}) = \nabla_{\mathbf{x}_0}\mathrm{J}^{\mathrm{p}}(\mathbf{x}_0^{\mathrm{b}}) \tag{28}$$

$$\mathbf{P}_k^{\mathrm{a}} = \mathbf{L}(k,0)\mathbf{P}_0^{\mathrm{a}}\mathbf{L}^{\mathrm{T}}(k,0). \tag{29}$$

These relations are obtained by simply applying $\mathbf{L}(k,0)$ to Eqs. (22), (23), (24), and (25).

The expressions for Eqs. (26) and (29) show that the optimal solution associated with the cost function Eq. (16) is on the trajectory initiated with $\hat{\mathbf{x}}_0$, while the associated analysis error covariance is propagated by the model. This is true whether the model is invertible or not (see appendix for an explanation).

(ii) *Additive property.* For the cost function Eq. (7), we process all the observations simultaneously. We may separate the observations into a few subsets. The 4D-Var process is performed for each subset as implemented by Jarvinnen *et al.* (1996). Here we want to prove that the final solution can be identical to the one obtained when all the observations are simultaneously processed.

We discuss the case in which the observations are divided into two subsets. Two cost functions associated with the observation subset are then

$$\mathrm{J}_1^{\mathrm{p}}(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})$$

$$+ \frac{1}{2}\sum_{i=1}^{l_1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_1(i)} - \mathbf{y}_{k_1(i)})^{\mathrm{T}}\mathbf{R}_{k_1(i)}^{-1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_1(i)} - \mathbf{y}_{k_1(i)}), \tag{30}$$

$$\mathrm{J}_2^{\mathrm{p}}(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^{\mathrm{T}}\hat{\mathbf{B}}_0^{-1}(\mathbf{x} - \hat{\mathbf{x}}_0)$$

$$+ \frac{1}{2}\sum_{i=1}^{l_2}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)})^{\mathrm{T}}\mathbf{R}_{k_2(i)}^{-1}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)}), \tag{31}$$

where $l = l_1 + l_2$, and $k_1(i)$ $(i = 1, \ldots, l_1)$ and $k_2(i)$ $(i = 1, \ldots, l_2)$ sum up to yield k $(k = 1, \ldots, l)$. The 4D-Var calculation is first carried out with J_1^p with respect to $\mathbf{x}_0$. The optimal solution $\hat{\mathbf{x}}_0$ is used as the background in J_2^p. $\hat{\mathbf{B}}_0$ is the analysis error covariance associated with $\hat{\mathbf{x}}_0$. Then the 4D-Var calculation is carried out with J_2^p. The final solution is denoted by $\hat{\hat{\mathbf{x}}}_0$.

What we need to prove is that $\hat{\hat{\mathbf{x}}}_0$ is the same solution as the one that directly minimizes Eq. (7). We may write J_1^p as

$$J_1^p(\mathbf{x}_0) = (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T \mathbf{H}_{0,l_1}(\mathbf{x}_0 - \hat{\mathbf{x}}_0) + J_1^p(\hat{\mathbf{x}}_0), \tag{32}$$

where $\mathbf{H}_{0,l_1}$ is the Hessian of the cost function Eq. (30). Equation (32) is actually the Taylor expansion of $J_1^p(\mathbf{x}_0)$ at $\hat{\mathbf{x}}_0$. Note that the first derivative is zero at $\hat{\mathbf{x}}_0$.

Since $\mathbf{H}_{0,l_1}^{-1}$ in Eq. (32) is the analysis error covariance of $\hat{\mathbf{x}}_0$, it is simply $\hat{\mathbf{B}}_0$ in Eq. (31). We then obtain

$$J^p(\mathbf{x}_0) = J_2^p(\mathbf{x}_0) + J_1^p(\mathbf{x}_0) - \tfrac{1}{2}(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T \hat{\mathbf{B}}_0^{-1}(\mathbf{x} - \hat{\mathbf{x}}_0) = J_2^p(\mathbf{x}_0) + J_1^p(\hat{\mathbf{x}}_0). \tag{33}$$

Since $J_1^p(\hat{\mathbf{x}}_0)$ is a constant, the minimization solution of $J^p(\mathbf{x}_0)$ is identical to that of $J_2^p(\mathbf{x}_0)$ when both problems are solved accurately.

(*b*) *Model with errors*

(i) *Consistent properties of the time interval.* When model errors are considered, the cost function in 4D-Var can be written as (Jazwinski 1970; Cohn 1997)

$$J^e = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^b)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^b) + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)$$
$$+ \frac{1}{2}\sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k. \tag{34}$$

This cost function implies that the model errors are time uncorrelated. The case where the correlation amongst model errors is taken into account has been investigated (see Zupanski 1997).

We have three choices for carrying out the minimization of the cost function Eq. (34), that is:

1. Carrying out constrained minimization J^e with respect to $(\mathbf{x}_0, \ldots, \mathbf{x}_l, \mathbf{w}_1, \ldots, \mathbf{w}_l)$, subject to Eq. (1). In this case the optimization is computationally prohibitive for this formulation in meteorological data assimilation, though it has been applied by using 'the representer reduction approximation' in oceanic data assimilation (e.g. Bennett 1992). This formulation is usually employed for theoretical analysis, and it has been exclusively used to prove the equivalence between Kalman smoothers and 4D-Var (Bryson and Ho 1975; Bennett and Budgell 1989; Ménard and Daley 1996; Zhu *et al.* 1999).

2. Carrying out unconstrained minimization of the following cost function (Cohn 1997)

$$J_1^e = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^b)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^b) + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)$$
$$+ \frac{1}{2}\sum_{k=1}^{l}\{\mathbf{x}_k - \mathbf{L}(k, k-1)\mathbf{x}_{k-1}\}^T \Gamma^T(k-1)\mathbf{Q}_k^{-1}\Gamma(k-1)\{\mathbf{x}_k - \mathbf{L}(k, k-1)\mathbf{x}_{k-1}\}, \tag{35}$$

with respect to $(\mathbf{x}_0, \ldots, \mathbf{x}_l)$. In this sense, we should not be considering 4D-Var as a method for searching for the optimal initial condition, but rather for searching for an optimal trajectory segment.

3. Carrying out unconstrained minimization of the following cost function (also see Zupanski 1997)

$$J_2^e = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^b)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^b)$$
$$+ \frac{1}{2}\sum_{k=1}^{l}\left\{\mathbf{h}_k\left(\mathbf{L}(k, 0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k, j)\Gamma(j-1)\mathbf{w}_j\right) - \mathbf{y}_k\right\}^T$$
$$\times \mathbf{R}_k^{-1}\left\{\mathbf{h}_k\left(\mathbf{L}(k, 0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k, j)\Gamma(j-1)\mathbf{w}_j\right) - \mathbf{y}_k\right\}$$
$$+ \frac{1}{2}\sum_{k=1}^{l}\mathbf{w}_k^T \mathbf{Q}_k^{-1}\mathbf{w}_k, \tag{36}$$

with respect to $\{\mathbf{x}_0, \mathbf{w}_1 \ldots, \mathbf{w}_l\}$. In the derivation of J_2^e, we have used the time integration solution of the difference equation (1),

$$\mathbf{x}_k = \mathbf{L}(k, 0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k, j)\Gamma(j-1)\mathbf{w}_j. \tag{37}$$

The merit of J_2^e deserves to be emphasized. We know that $\mathbf{w}_k$ are r-vectors. Practically, r may be smaller than the dimension n of $\mathbf{x}_k$. Specifically, J_2^e provides us with tremendous flexibility in approximately representing model errors with a small r. Such an approximation has been implemented by Zupanski (1997). In the variational continuous-assimilation technique due to Derber (1989), $\mathbf{w}_k$ is simply considered to be independent of k, that is, the time variable is fixed. In the following, we use J_2^e, and thus the derivation is computational and algorithm oriented.

Since we treat $\{\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l\}$ jointly, let us introduce an augmented $n + lr$ vector $\mathbf{Z}_0^T = (\mathbf{x}_0^T, \mathbf{w}_1^T, \ldots, \mathbf{w}_l^T)$. Then J_2^e is transformed to

$$J_2^e = \frac{1}{2}(\mathbf{Z}_0 - \mathbf{Z}_0^b)^T \mathbf{B}_{Z_0}^{-1}(\mathbf{Z}_0 - \mathbf{Z}_0^b) + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{C}_k\mathbf{Z}_0 - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{C}_k\mathbf{Z}_0 - \mathbf{y}_k), \tag{38}$$

where $(\mathbf{Z}_0^b)^T = \{(\mathbf{x}_0^b)^T, 0, \ldots, 0\}$, $\mathbf{B}_{Z_0}$ is a $(1 + l) \times (1 + l)$ block diagonal matrix whose diagonal blocks are $\mathbf{B}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_l$ (thus $\mathbf{B}_{Z_0}$ is an $(n + rl) \times (n + rl)$ matrix),

and

$$\mathbf{C}_k = \{\mathbf{L}(k, 0), \mathbf{L}(k, 1)\Gamma(0), \ldots, \mathbf{L}(k, k)\Gamma(k-1), 0, \ldots, 0\} \tag{39}$$

are $n \times (n + rl)$ matrices.

For the case of the perfect model J_2^e reverts to an identical formulation to J^p. As for J^p, we obtain the optimal solution

$$\hat{\mathbf{Z}}_0 = \mathbf{Z}_0^b + \mathbf{H}_{\mathbf{Z}_0}^{-1} \nabla_{\mathbf{Z}_0} J_2^e(\mathbf{Z}_0^b), \tag{40}$$

where the Hessian matrix is

$$\mathbf{H}_{\mathbf{Z}_0} = \mathbf{B}_{Z_0}^{-1} + \sum_{k=1}^{l} \mathbf{C}_k^T \mathbf{h}_k^T \mathbf{R}_k^{-1} \mathbf{h}_k \mathbf{C}_k \tag{41}$$

and the gradient of J_2^e at $\mathbf{Z}_0^b$

$$\begin{aligned} \nabla_{\mathbf{Z}_0} J_2^e(\mathbf{Z}_0^b) &= \sum_{k=1}^{l} \mathbf{C}_k^T \mathbf{h}_k^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{C}_k \mathbf{Z}_0^b - \mathbf{y}_k) \\ &= \sum_{k=1}^{l} \mathbf{C}_k^T \mathbf{h}_k^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{L}(k, 0) \mathbf{x}_0^b - \mathbf{y}_k). \end{aligned} \tag{42}$$

Also, we can calculate that the error covariance of $\hat{\mathbf{Z}}_0$ is the inverse of the Hessian matrix, that is

$$\mathbf{P}_{\mathbf{Z}_0}^a = \mathbf{H}_{\mathbf{Z}_0}^{-1}. \tag{43}$$

Now we turn to proving that every model state on the trajectory determined by $\hat{\mathbf{Z}}_0 = (\hat{\mathbf{x}}_0^T, \hat{\mathbf{w}}_1^T, \ldots, \hat{\mathbf{w}}_l^T)$ minimizes Eq. (36). From Eq. (37), the model state on this trajectory is

$$\hat{\mathbf{x}}_k = \mathbf{L}(k, 0)\hat{\mathbf{x}}_0 + \sum_{j=1}^{k} \mathbf{L}(k, j)\Gamma(j-1)\hat{\mathbf{w}}_j = \mathbf{C}_k \hat{\mathbf{Z}}_0. \tag{44}$$

That is, we need to prove that $\hat{\mathbf{Z}}_k = \{\mathbf{L}(k, 0)\hat{\mathbf{x}}_0^T, \hat{\mathbf{w}}_1^T, \ldots, \hat{\mathbf{w}}_l^T\}$ also minimizes Eq. (36). This is obvious when considering $\mathbf{Z}_k = \mathbf{A}(k, 0)\mathbf{Z}_0$ and making a comparison with the case of the perfect model, and we further have

$$\mathbf{A}^T(k, 0)\mathbf{H}_{\mathbf{Z}_k}\mathbf{A}(k, 0) = \mathbf{H}_{\mathbf{Z}_0} \tag{45}$$

$$\mathbf{P}_{\mathbf{Z}_k}^a = \mathbf{A}(k, 0)\mathbf{P}_{\mathbf{Z}_0}^a \mathbf{A}^T(k, 0). \tag{46}$$

Here $\mathbf{A}(k, 0)$ is an $(n + rl) \times (n + rl)$ matrix. Its first n rows are $\mathbf{C}(k, 0)$, and the diagonal entries are 1 with the non-diagonal entries being zero for all the other rows.

(ii) *Additive property.* 4D-Var with models where errors are present also retains the additive property similar to that for perfect models. We define two cost functions

$$\begin{aligned} J^e_{2,1}(\mathbf{Z}_0) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b_0)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}^b_0) \\ &\quad + \frac{1}{2}\sum_{i=1}^{l_1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_1(i)} - \mathbf{y}_{k_1(i)})^T \mathbf{R}^{-1}_{k_1(i)}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_1(i)} - \mathbf{y}_{k_1(i)}) \\ &\quad + \sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1}\mathbf{w}_k, \end{aligned} \tag{47}$$

$$\begin{aligned} J^e_{2,2}(\mathbf{Z}_0) &= \frac{1}{2}(\mathbf{Z}_0 - \hat{\mathbf{Z}}_0)^T \hat{\mathbf{B}}^{-1}_{\mathbf{Z}_0}(\mathbf{Z} - \hat{\mathbf{Z}}_0) \\ &\quad + \frac{1}{2}\sum_{i=1}^{l_2}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)})^T \mathbf{R}^{-1}_{k_2(i)}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)}). \end{aligned} \tag{48}$$

Suppose that $\hat{\mathbf{Z}}_0$ is the minimization solution of $J^e_{2,1}(\mathbf{Z}_0)$ and $\hat{\mathbf{B}}^{-1}_{\mathbf{Z}_0}$ is taken as $\mathbf{H}_{\mathbf{Z}_0}$. The Taylor expansion yields

$$J^e_{2,1}(\mathbf{Z}_0) = J^e_{2,1}(\hat{\mathbf{Z}}_0) + \tfrac{1}{2}(\mathbf{Z} - \hat{\mathbf{Z}}_0)^T \hat{\mathbf{B}}^{-1}_{\mathbf{Z}}(\mathbf{Z} - \hat{\mathbf{Z}}_0). \tag{49}$$

Combining Eqs. (48) and (49), we obtain

$$J^e_2(\mathbf{Z}_0) = J^e_{2,2}(\mathbf{Z}_0) + J^e_{2,1}(\hat{\mathbf{Z}}_0). \tag{50}$$

Thus, the minimization solution of $J^e_{2,2}(\mathbf{Z})$ is identical to that of $J^e_2(\mathbf{Z})$, where all the observations are processed simultaneously.

4. Optimality of four-dimensional variational data assimilation

(a) *Consistent and additive properties of optimality*

(i) *Perfect model.* For a perfect model, it has been shown (Lorenc 1986) that the joint conditional probability density is

$$\begin{aligned} &p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l) \\ &\quad = c \exp\left(-\frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b_0)^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}^b_0) - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k)^T \mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k)\right), \end{aligned} \tag{51}$$

where c is a constant. Thus, maximizing $p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l)$ with respect to $\{\mathbf{x}_0, \ldots, \mathbf{x}_l\}$ is equivalent to minimizing Eq. (7). The minimum solution of the cost function Eq. (7) is the *joint* maximum likelihood (Bayesian) estimate or the most probable estimate.

Now we are concerned with the optimal solution with respect to $\mathbf{x}_k$ $(0 \le k \le l)$ only, that is, the solution that maximizes the marginal probability density

$$p(\mathbf{x}_k | \mathbf{y}_1, \ldots, \mathbf{y}_l), \tag{52}$$

rather than the joint probability density.

We first consider the case for $k = 0$. Using Bayes' rule, we get

$$p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l) = \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_l|\mathbf{x}_0)p(\mathbf{x}_0)}{p(\mathbf{y}_1, \ldots, \mathbf{y}_l)}. \tag{53}$$

In view of Eq. (4), we get

$$\mathbf{y}_i - \mathbf{h}_i\mathbf{L}(i, 0)\mathbf{x}_0 = \mathbf{v}_i \quad i = 1, \ldots, l. \tag{54}$$

Thus, with $\mathbf{x}_0$ given and with $\{\mathbf{x}_1 \ldots, \mathbf{x}_l\}$ given by $\mathbf{x}_k = \mathbf{L}(k, 0)\mathbf{x}_0$, $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$ are independent. We then have

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_l|\mathbf{x}_0) &= c'p\{\mathbf{y}_1|\mathbf{L}(1, 0)\mathbf{x}_0)\} \ldots p\{(\mathbf{y}_l|\mathbf{L}(l, 0)\mathbf{x}_0\} \\ &= p_{\mathbf{v}_1}\{\mathbf{y}_1 - \mathbf{h}_1\mathbf{L}(1, 0)\mathbf{x}_0\} \ldots p_{\mathbf{v}_l}\{\mathbf{y}_l - \mathbf{h}_l\mathbf{L}(l, 0)\mathbf{x}_0\}, \end{aligned} \tag{55}$$

where c' is a constant. Since $\mathbf{x}_0$ and $\mathbf{v}_i$ are Gaussian, Eq. (53) becomes

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l) = c \exp\Bigg(&-\frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}) \\ &- \frac{1}{2}\sum_{i=1}^{l}\{\mathbf{h}_i\mathbf{L}(i, 0)\mathbf{x}_0 - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{R}_i^{-1}\{\mathbf{h}_i\mathbf{L}(i, 0)\mathbf{x}_0 - \mathbf{y}_i\}\Bigg), \end{aligned} \tag{56}$$

where c is a constant depending only on c' and $p(\mathbf{y}_1, \ldots, \mathbf{y}_l)$. Thus, maximizing $p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l)$ turns out to be equivalent to minimizing Eq. (8). Due to the consistency property of 4D-Var, maximizing $p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l)$ is equivalent to maximizing $p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l|\mathbf{y}_1, \ldots, \mathbf{y}_l)$.

Since $\mathbf{x}_k = \mathbf{L}(k, 0)\mathbf{x}_0$, we have (Jazwinski 1970, Theorem 2.7)

$$\|\mathbf{L}(k, 0)\|p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l) = p(\mathbf{x}_k|\mathbf{y}_1, \ldots, \mathbf{y}_l), \tag{57}$$

where $\|\mathbf{L}(k, 0)\|$ is the absolute value of the determinant of $\mathbf{L}(k, 0)$. Thus, maximizing $p(\mathbf{x}_k|\mathbf{y}_1, \ldots, \mathbf{y}_l)$ is equivalent to maximizing $p(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_l)$.

We then draw the conclusion that the 4D-Var solution is optimal not only with respect to the model trajectory consisting of $\{\mathbf{x}_0, \ldots, \mathbf{x}_l\}$, but also with respect to $\mathbf{x}_k$ at a single observation time.

(ii) *Model with errors.* Here we first consider maximizing

$$p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l|\mathbf{y}_1, \ldots, \mathbf{y}_l) \tag{58}$$

with respect to $\{\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l\}$. Similar to Jazwinski (1970, pp. 153–155), we shall show that the 4D-Var solution minimizing the cost function Eq. (36) is identical to the maximum likelihood estimate.

Using Bayes' rule, we write the density in Eq. (58) as

$$\begin{aligned} &p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l|\mathbf{y}_1, \ldots, \mathbf{y}_l) \\ &\quad = \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_l|\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l)p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l)}{p(\mathbf{y}_1, \ldots, \mathbf{y}_l)} \\ &\quad = \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_l|\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l)p(\mathbf{x}_0|\mathbf{w}_1, \ldots, \mathbf{w}_l)p(\mathbf{w}_1, \ldots, \mathbf{w}_l)}{p(\mathbf{y}_1, \ldots, \mathbf{y}_l)}. \end{aligned} \tag{59}$$

With $\{\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l\}$ given, $\mathbf{x}_k$ is then determined by Eq. (37). In view of Eq. (4), with $\mathbf{x}_k$ given, $\mathbf{y}_k$ are Gaussian and independent, since $\mathbf{v}_k$ are white. We thus have

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l) &= p(\mathbf{y}_1 | \mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l) \ldots p(\mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l) \\ &= p(\mathbf{y}_1 | \mathbf{x}_1) \ldots p(\mathbf{y}_l | \mathbf{x}_l) \\ &= p_{\mathbf{v}_1}\left\{\mathbf{y}_1 - \mathbf{h}_1\left(\mathbf{L}(k, 0)x_0 + \sum_{j=1}^{1} \mathbf{L}(1, j)\Gamma(j-1)\mathbf{w}_j\right)\right\} \\ &\quad \ldots p_{\mathbf{v}_l}\left\{\mathbf{y}_l - \mathbf{h}_l\left(\mathbf{L}(k, 0)x_0 + \sum_{j=1}^{l} \mathbf{L}(l, j)\Gamma(j-1)\mathbf{w}_j\right)\right\}. \end{aligned} \tag{60}$$

Since $\mathbf{x}_0$ is independent of $\{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$, we obtain

$$p(\mathbf{x}_0 | \mathbf{w}_1, \ldots, \mathbf{w}_l) = p_{\mathbf{x}_0}(\mathbf{x}_0). \tag{61}$$

With $\{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$ independent, we get

$$p(\mathbf{w}_1, \ldots, \mathbf{w}_l) = p_{\mathbf{w}_1}(\mathbf{w}_1) \ldots p_{\mathbf{w}_l}(\mathbf{w}_l). \tag{62}$$

Substitution of Eqs. (60), (61) and (62) into Eq. (59) yields

$$\begin{aligned} p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l) &= c' \exp\left[-\frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})^{\mathrm{T}} \mathbf{B}_0^{-1} (\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}})\right. \\ &\quad - \frac{1}{2}\sum_{k=1}^{l}\left\{\mathbf{h}_k\left(\mathbf{L}(k, 0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k, j)\Gamma(j-1)\mathbf{w}_j\right) - \mathbf{y}_k\right\}^{\mathrm{T}} \\ &\quad \times \mathbf{R}_k^{-1}\left\{\mathbf{h}_k\left(\mathbf{L}(k, 0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k, j)\Gamma(j-1)\mathbf{w}_j\right) - \mathbf{y}_k\right\} \\ &\quad \left. - \frac{1}{2}\sum_{k=1}^{l} \mathbf{w}_k^{\mathrm{T}} \mathbf{Q}_k^{-1} \mathbf{w}_k\right], \end{aligned} \tag{63}$$

where c' only depends on $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$, and is independent of $\{\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l\}$. Thus, maximizing $p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l)$ is equivalent to minimizing the cost function Eq. (36).

From Eq. (49), we can conclude that $p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{Y}_l)$ is Gaussian, and thus $p(\mathbf{x}_0 | \mathbf{Y}_l)$, $p(\mathbf{w}_1 | \mathbf{Y}_l)$, ... $p(\mathbf{w}_l | \mathbf{Y}_l)$ are also Gaussian. The minimum solution of Eq. (36), $\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_l$, provides the maximum likelihood estimates, that is, $\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_l$ maximize $p(\mathbf{x}_0 | \mathbf{Y}_l)$, $p(\mathbf{w}_1 | \mathbf{Y}_l)$, ... $p(\mathbf{w}_l | \mathbf{Y}_l)$, respectively.

Now we turn to the optimal estimation with respect to $\{\mathbf{x}_k, \mathbf{w}_1, \ldots, \mathbf{w}_l\}$, $0 \leq k \leq l$, that is, the estimate maximizes the conditional density

$$p(\mathbf{x}_k, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l). \tag{64}$$

Let $\mathbf{Z}_k^{\mathrm{T}} = (\mathbf{x}_k, \mathbf{w}_1, \ldots, \mathbf{w}_l)^{\mathrm{T}}$. Since we have the transformation

$$\mathbf{Z}_k^{\mathrm{T}} = \mathbf{A}(k, 0)\mathbf{Z}_0^{\mathrm{T}}, \tag{65}$$

where $\mathbf{A}(k, 0)$ is defined in section 3(b). Then we have (Jazwinski 1970, Theorem 2.7)

$$\|\mathbf{A}\| p(\mathbf{x}_k, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l) = p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l), \tag{66}$$

where $\|\mathbf{A}\|$ is the absolute value of the determinant of $\mathbf{A}$. Maximizing $p(\mathbf{x}_k, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l)$ is thus equivalent to maximizing $p(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_l | \mathbf{y}_1, \ldots, \mathbf{y}_l)$.

We thus conclude that the 4D-Var solution is optimal, not only with respect to the model trajectory consisting of $(\mathbf{x}_0, \ldots, \mathbf{x}_l)$, but also with respect to $\mathbf{x}_k$ at a single observation time even though the model error is taken into consideration. We call this property the *consistent optimality*.

When observations are separated into subsets, 4D-Var is carried out in the way described in section 3. From the additive property, we know that the final 4D-Var solution is optimal as all the observations are processed simultaneously.

(*b*) *Optimal transferability property*

We further present an optimal property of 4D-Var, which we call the transferable optimality property. This property is intimately related to the equivalence between 4D-Var and Kalman smoothers. Here we consider a time window with observations

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_l, \mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m}\}. \tag{67}$$

The observations are split into two batches in time, namely,

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m}\}. \tag{68}$$

We solve two 4D-Var problems: the first one with observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$, yielding the solution $\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_l\}$; the second one with $\{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m}\}$, yielding $\{\hat{\mathbf{x}}_l, \ldots, \hat{\mathbf{x}}_{l+m}\}$. The background of the second 4D-Var is $\hat{\mathbf{x}}_l$, that is, the solution of the first 4D-Var of the first batch. Then the final 4D-Var solution yielded by the second batch is optimal as in the 4D-Var where all data in the two batches are processed simultaneously, namely, $\hat{\mathbf{x}}_k$ maximizes $p(\mathbf{x}_k | Y_{l+m})$ for $(l \le k \le l + m)$. This property shows that information from one batch can be fully transferred to the next one via the background term defined by $\{\hat{\mathbf{x}}_l\}$, that is, the 4D-Var solution at the end of the time window of the previous batch.

For a perfect model in the linear context, 4D-Var preserves this property without any additional assumptions. Correspondingly, we can prove this property by a direct examination of the 4D-Var solution. For a model with errors, we can prove this property by invoking the property of Gaussian distributions. The proof is as follows.

Using Bayes' rule repeatedly, we can have

$$\begin{aligned} & p(\mathbf{x}_l, \ldots, \mathbf{x}_{l+m} | \mathbf{Y}_{l+m}) \\ & \quad = \frac{p(\mathbf{x}_l, \ldots, \mathbf{x}_{l+m}, \mathbf{Y}_l, \mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m})}{p(\mathbf{Y}_{l+m})} \\ & \quad = \frac{p(\mathbf{Y}_l)}{p(\mathbf{Y}_{l+m})} p(\mathbf{x}_l | \mathbf{Y}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m}) \\ & \qquad \times p(\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m} | \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}, \mathbf{x}_l, \mathbf{Y}_l) p(\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m} | \mathbf{x}_l, \mathbf{Y}_l). \end{aligned} \tag{69}$$

Since the sequence $\mathbf{x}_k$ $(k = 1, \ldots, l + m)$ is Markov, we have

$$p(\mathbf{x}_l | \mathbf{Y}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m}) = p(\mathbf{x}_l | \mathbf{Y}_l). \tag{70}$$

In view of Eq. (4), we get

$$\begin{aligned} & p(\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m} | \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}, \mathbf{x}_l, \mathbf{Y}_l) \\ & \quad = p(\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{l+m} | \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}) \\ & \quad = p_{\mathbf{v}_{l+1}}(\mathbf{y}_{l+1} - \mathbf{h}_{l+1}\mathbf{x}_{l+1}) \ldots p_{\mathbf{v}_{l+m}}(\mathbf{y}_{l+m} - \mathbf{h}_{l+m}\mathbf{x}_{l+m}). \end{aligned} \tag{71}$$

In view of Eqs. (4) and (37), $\mathbf{w}_k$ is independent of $\mathbf{Y}_l$ for $k > l$. Further, since the sequence $\mathbf{x}_k$ is Markov, we obtain

$$\begin{aligned} p(\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}|\mathbf{x}_l, \mathbf{Y}_l) &= p(\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+m}|\mathbf{x}_l) \\ &= p(\mathbf{x}_{l+m}|\mathbf{x}_{l+m-1}) \ldots p(\mathbf{x}_{l+1}|\mathbf{x}_l) \\ &= p_{\mathbf{w}_{l+1}}\{\mathbf{x}_{l+1} - \mathbf{L}(l+1, l)\mathbf{x}_l\} \ldots \\ &\quad \ldots p_{\mathbf{w}_{l+m}}\{\mathbf{x}_{l+m} - \mathbf{L}(l+m, l+m-1)\mathbf{x}_{l+m-1}\}. \end{aligned} \tag{72}$$

Substituting Eqs. (70), (71) and (72) into Eq. (69), we thus obtain that maximizing $p(\mathbf{x}_l, \ldots, \mathbf{x}_{l+m}|\mathbf{Y}_{l+m})$ is equivalent to minimizing the cost function

$$\begin{aligned} \mathrm{J} = &\frac{1}{2}(\mathbf{x}_l - \hat{\mathbf{x}}_l)^{\mathrm{T}}(\mathbf{P}_l^{\mathrm{a}})^{-1}(\mathbf{x}_l - \hat{\mathbf{x}}_l) + \frac{1}{2}\sum_{k=l+1}^{l+m}\{\mathbf{h}_k\mathbf{L}(k, l)\mathbf{x}_l - \mathbf{y}_k)^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{L}(k, l)\mathbf{x}_l - \mathbf{y}_k\} \\ &+ \frac{1}{2}\sum_{k=l+1}^{l+m}\mathbf{w}_k^{\mathrm{T}}\mathbf{Q}_k^{-1}\mathbf{w}_k, \end{aligned} \tag{73}$$

subject to constraints in Eq. (1), where $\hat{\mathbf{x}}_l$ and $\mathbf{P}_l^{\mathrm{a}}$ are provided by the 4D-Var process over the first batch.

5. Relationship between four-dimensional variational data assimilation and the Kalman filter/smoother

(a) *The Kalman filter*

It has long been known that for a perfect, linear model and linear observation operators, both 4D-Var and the Kalman filter (Kalman 1960) yield the same values for the model variables at the end of the 4D-Var assimilation period (also see Lorenc 1986; Ghil 1989). The previous section has shown that 4D-Var provides a maximum likelihood solution at the end of the time interval for both perfect and imperfect models, identical to the one yielded by the Kalman filter.

This equivalence can also be verified by directly comparing the 4D-Var and Kalman filter solutions. For one iteration of the Kalman filter, we have at the $(k-1)$th step

$$\mathbf{x}_{k-1} \sim \mathrm{N}(\hat{\mathbf{x}}_{k-1}, \mathbf{P}_{k-1}^{\mathrm{a}}). \tag{74}$$

We define a cost function corresponding to Eq. (73)

$$\begin{aligned} \mathrm{J}_k = &\tfrac{1}{2}(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1})^{\mathrm{T}}\mathbf{B}_{k-1}^{-1}(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1}) \\ &+ \tfrac{1}{2}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k)^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k) + \tfrac{1}{2}\mathbf{w}_k^{\mathrm{T}}\mathbf{Q}_k^{-1}\mathbf{w}_k. \end{aligned} \tag{75}$$

The optimal solution $\hat{\mathbf{x}}_k$ of Eq. (75) is identical to the Kalman filter solution (Bryson and Ho 1975, pp. 391–393). Based on the transferable property of 4D-Var, the equivalence between 4D-Var and the Kalman filter becomes obvious.

(b) *The Kalman smoother*

Generally, there are three types of smoothing (Anderson and Moore 1979, p. 166). Fixed-interval smoothing is concerned with obtaining optimal estimate $\hat{\mathbf{x}}_k$ using observations $\mathbf{Y}_l$ for fixed l and all k in the interval $0 \le k \le l$. Fixed-lag smoothing is concerned with obtaining optimal estimate $\hat{\mathbf{x}}_{l-n}$ using observations $\mathbf{Y}_l$ for all l and fixed n. Fixed-point smoothing is concerned with obtaining optimal estimate $\hat{\mathbf{x}}_j$ using $\mathbf{Y}_l$ for

fixed j and all l, where $0 \leq j \leq l$. Fixed-point smoothing does not seem to be applicable in meteorology, thus it will not be addressed.

In the linear context, 4D-Var is theoretically a perfect smoother, since the 4D-Var solution is optimal not only with respect to the model trajectory consisting of $(\mathbf{x}_0, \ldots, \mathbf{x}_l)$, but also with respect to $\mathbf{x}_k$ at a single observation time level. One single evaluation of 4D-Var yields all optimal estimates $(\hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_l)$.

In practice, however, 4D-Var is always performed over a limited time interval. The fixed-lag smoothing is required to be implemented for any l. For large l, 4D-Var cannot be directly applied. This is the case when we are producing a consistent and long sequence of assimilation data, or climatological data. Producing a set of time-consistent climatological data is the purpose of all re-analysis projects (e.g. Schubert and Rood 1995; Gibson *et al.* 1996; Kalnay *et al.* 1996). This issue has been discussed in depth by Cohn *et al.* (1994) and Todling *et al.* (1998). They note that retrospective data assimilation based on fixed-lag smoothing should be the ultimate goal of re-analysis efforts. Thus, 4D-Var is practically not equivalent to the fixed-lag smoother.

(*c*) *A four-dimensional variational fixed-lag smoother*

Although 4D-Var is practically not equivalent to the fixed-lag smoother, we can use 4D-Var to construct a smoother that is equivalent to the fixed-lag smoother proposed by Cohn *et al.* (1994). In what follows, we propose such an approach.

This approach is based on a batch-processing approach or cycling 4D-Var. The batch-processing approach consists of separating the observations into batches in time:

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{2l}\}, \ldots, \{\mathbf{y}_{(j-1)l+1}, \ldots, \mathbf{y}_{jl}\}, \ldots, \tag{76}$$

and solving the standard 4D-Var problem separately for each batch of observations (Jazwinski 1970).

Due to the optimality of the transferability property of 4D-Var, it immediately follows that the optimal solution $\hat{\mathbf{x}}_{(j-1)l+k}$ in the jth batch maximizes

$$p(\mathbf{x}_{(j-1)l+k} | \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_j) \tag{77}$$

where $1 \leq k \leq l$, and $\mathbf{Y}_j = \{\mathbf{y}_{(j-1)l+1}, \ldots, \mathbf{y}_{jl}\}$.

This transferable optimality property of the batch processing approach allows us to construct a smoother which is equivalent to the fixed-lag Kalman smoother proposed by Cohn *et al.* (1994). The fixed-lag Kalman smoother seeks the optimal $\hat{\mathbf{x}}_k$ that maximizes $p(\mathbf{x}_k | \mathbf{y}_1, \ldots, y_k, \ldots, y_{k+n})$ for all k with n fixed. The constructive procedure is outlined as follows:

1. We separate the observations into a sequence of batches,

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_{2l}\}, \ldots, \tag{78}$$

   and proceed to solve the standard 4D-Var problem separately for each batch of observations. The solution at the end of one batch is used as the background for the subsequent batch.

2. Since assimilated analyses are not required at every time level of observations, the time interval between required assimilated analyses, for instance, is d times as large as the observation interval. We can construct new sequences of batches in time,

$$\{\mathbf{y}_{1+md}, \ldots, \mathbf{y}_{l+md}\}, \{\mathbf{y}_{l+md+1}, \ldots, \mathbf{y}_{2l+md}\}, \ldots, \tag{79}$$

where $m = 1, \ldots, l/d \leq l$. Here we have assumed that the time interval with each batch is $M = l/d$ times as large as the time interval between required assimilated analyses.

3. Using these $M + 1$ assimilated analysis sequences, we can construct a time-consistent climatological data sequence by picking one single assimilation data point in each batch, in order that the fixed n $(1 \leq n \leq l)$ future time levels of observations are used for every assimilated analysis $\hat{\mathbf{x}}_k$. Thus, $\hat{\mathbf{x}}_k$ maximizes $p(\mathbf{x}_k|\mathbf{y}_1, \ldots, y_k, \ldots, y_{k+n})$ for all k.

The assimilated data sequence produced by one sequence of batches defined in Eq. (76) is not consistent in time. This actually causes a type of time discontinuity (Cohn *et al.* 1994). One may think of separating the observations in a different manner. There is an estimation principle which states that all observations must be used only once for the estimation of one variable. To comply with this principle, such batch processing appears to be the most efficient algorithm.

This 4D-Var fixed-lag smoother requires M-batch processing. M is usually one. This occurs, for instance, if the 4D-Var time window length is six hours and we require assimilation data for every six hours. However, even though $M > 1$, the computational overhead must not necessarily increase M fold. Although M-batch processing is carried out independently, the information generated by one batch processing can be used by another batch processing. This observation leads to two techniques which may be used jointly to accelerate the minimization procedure, and thus reduce the computational overhead.

First, the 4D-Var in the $(m-1)$th sequence provides the initial guess to the 4D-Var in the m-th sequence. This technique is the same as the one used in quasi-continuous variational data assimilation due to Jarvinnen *et al.* (1996). We note that in this case the initial guess should be distinguished from the background field. The background field *must* be provided by the 4D-Var in the previous batch in the same sequence in order to ensure that all observations are used only once for the estimation of each assimilated analysis. Second, the 4D-Var in the $(m-1)$th sequence provides the Hessian information for the 4D-Var in the m-th sequence. This Hessian information can be used to precondition the minimization procedure, or even to start directly the 4D-Var procedure (Courtier *et al.* 1994; Li *et al.* 2000).

6. Estimation of error covariances

The equivalence of 4D-Var to the Kalman filter/smoother hinges on the availability of analysis error covariances. 4D-Var itself does not involve generating the time evolution of covariances. However, the formulation presented above allows the evaluation of the error covariances in the framework of 4D-Var, and the method is essentially identical to that used for the Kalman filter and smoother.

We have shown that the analysis error covariance is the inverse of the Hessian of the cost function for both perfect models and imperfect models (see Eqs. (15) and (43)). This relationship provides us with possibilities for evaluating the analysis error covariance (also see Fisher and Courtier 1995). In fact, the inverse of the Hessian of 3D-Var is included in the Kalman gain matrix, and also has to be explicitly computed in the Kalman filter.

Sub-optimal techniques used in the Kalman filter/smoother are directly applicable to cycling 4D-Var. One method consists of computing directly the approximate of the inverse of the Hessian. Since the adjoint model is usually available, the eigenvalues and

eigenvectors can be computed efficiently (e.g. Molteni and Palmer 1993). A powerful method is thus to perform the singular vector decomposition of the Hessian, and then compute the inverse of the Hessian. Assume that the rank of the Hessian is $l \leq n$, and the eigenvalues and eigenvectors are σ_i^2 and $\mathbf{v}_i$ $(i = 0, \ldots, l)$. The analysis error covariance is

$$\mathbf{P}_{\mathbf{x}_0}^{\mathrm{a}} = \mathbf{V}\Sigma^{-1}\mathbf{V}^{\mathrm{T}}, \tag{80}$$

where Σ^{-1} is an $l \times l$ diagonal matrix with the diagonal entries of σ_i^{-2}, and $\mathbf{V}$ an $n \times l$ matrix with its columns of $\mathbf{v}_i$. For a sub-optimal technique, $\mathbf{V}$ contains only singular vectors associated with the dominant large singular values.

An alternative is to estimate the inverse of the Hessian approximately during the minimization process. Minimization algorithms usually provide an estimate of the inverse of the Hessian, such as quasi-Newton types of minimization algorithms. This implies the possibility of estimating the analysis error covariance matrix during the minimization procedure (Fisher and Courtier 1995).

When the above mentioned algorithms are applied to the case of imperfect models, the concern is the dimension of the Hessian $\mathbf{H}_{Z_0}$. How to use a limited number of parameters to represent model errors still constitutes a challenge, though there have been efforts in this direction such as Zupanski (1997). In this sense, sequential methods possess advantages, since sequential methods implement the computation once at one single observational time level rather than several time levels being processed jointly as in 4D-Var.

The above techniques are only concerned with computing the analysis covariance associated with $\mathbf{x}_0$ at the initial time. For $\mathbf{x}_k$, we must propagate the analysis covariance in terms of Eqs. (29) and (46), which is also required by the Kalman filter/smoother. To do so, sub-optimal methods are usually necessary. Examples are simplified Kalman filters (Fisher 1998) that are related to the partial eigen-decomposition method (Todling *et al.* 1998), and the ensemble method due to Evensen (1994).

When we resort to sub-optimal algorithms, an advantage of cycling 4D-Var is that error covariances are implicitly and accurately evolved within the 4D-Var time window. The implicit evolution of error covariances also results in computational efficiency.

7. Summary and discussion

We showed that the 4D-Var solution was optimal not only with respect to the model trajectory segment over the assimilation time interval but also with respect to any model state at a single observation time, irrespective of whether the model was perfect or not. In the batch-processing method, the information in 4D-Var was fully transferred from one batch to the next by the background term. As in the Kalman filter, 4D-Var may also fully extract information from observations prior to the time interval over which 4D-Var is currently being performed. Also, 4D-Var allows the processing of observations in subsets, while the final solution is optimal as all observations are processed simultaneously. It should be emphasized that these conclusions were true only for a linear model and linear observation operators. A major assumption in all the derivations in this study was that the distribution of errors was Gaussian. In a nonlinear system, the assumption of Gaussian distribution may not be applicable. Thus, for a nonlinear system, the solution generally does not possess such a property.

From the optimality of 4D-Var, the equivalence between 4D-Var and the Kalman filter/smoother became evident. 4D-Var is a perfect fixed-interval smoother in both

practical and theoretical senses in the linear context, even for models with errors. For the fixed-lag Kalman smoother, there is no direct equivalent relationship. However, we proposed a 4D-Var fixed-lag smoother. This smoother turned out to be equivalent to the fixed-lag Kalman smoother proposed by Cohn *et al.* (1994) in the linear context. Due to the computational efficiency of 4D-Var, this multi-batch processing 4D-Var smoother may hold promise for practical use.

Since a 4D-Var solution is optimal at any single observation time, and the information is fully transferred from one batch to the next by the background term, the variational method and sequential method can be used *interchangeably*. Difficulties usually arise when 4D-Var is used to produce a consistent and long sequence of assimilation data, which is the major purpose of data re-analysis (Schubert and Rood 1995; Gibson *et al.* 1996; Kalnay *et al.* 1996). Practically, 4D-Var can only be carried out for a limited time interval. On the other hand, the fixed-lag Kalman smoother, as a sequential method, is naturally applicable to this task. With the 4D-Var fixed-lag smoother that we constructed, 4D-Var can also produce a consistent and long sequence of climatological data.

The optimality properties of 4D-Var seem to lead to the conclusion that 4D-Var provides an assimilation identical to 3D-Var at the end of the time window, as the smoother provides an assimilation identical to the filter. However, we emphasize that approximations in computing error covariances may result in substantial differences among these methods.

The equivalence of 4D-Var to the Kalman filter/smoother hinges on two prerequisites: all statistics of the error covariances used in 4D-Var should be identical to those used in the Kalman filter/smoother; 4D-Var reaches the exact minimum solution of the problem. Unfortunately, this is not the case in practice. Sub-optimal use of covariances will be inevitable. 4D-Var possesses an advantage in that it only approximates error covariances at the beginning or end of the time window, while the propagation of error covariances in the time window is implicit and thus accurate. In sequential methods, approximation is made for each time step. However, the minimization process in 4D-Var is generally terminated prior to the minimal solution being obtained. Such termination may not cause a serious problem for 4D-Var used for NWP, but it constitutes a serious concern when 4D-Var is used for producing climatological data, as in the proposed multi-processing 4D-Var smoother.

A common challenge for both the sequential and variational methods is how to deal with model errors. A serious difficulty consists of how to define the statistics of model errors, although model bias seems manageable (e.g. Dee and da Silva 1998; Derber 1989). In 4D-Var, it is necessary to use only a limited number of parameters to represent model errors, as shown in section 3 (also see Bennett 1992; Zupanski 1997). Further, time correlation of model errors may necessarily be considered (Cohn and Dee 1988).

We have emphasized above the optimality of 4D-Var, but have made no attempt to determine whether the sequential or variational method performs better. In practice, the performance of each method strongly depends on model performance, quantity of observations, and use of assimilated data. Moreover, approximations in computing error covariances are inevitable and object oriented.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their insightful comments and suggestions, which substantially improved this paper. Z. Li also thanks Dr Todling at the Data Assimilation Office for his welcome suggestions. We acknowledge the support from

the National Science Foundation grant number ATM-9731472 managed by Dr Pamela Stephens whom we would like to thank for her support.

APPENDIX

Invertibility of models and 4D-Var

Let us consider the contribution of one observation, $\mathbf{y}_k$, to the optimal solution, $\hat{\mathbf{x}}_0$, and the reduction in the analysis error variance. As such, the Hessian (see Eq. (12)) is

$$\mathbf{H}_{0,l} = \mathbf{B}_0^{-1} + \mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k, 0). \tag{A.1}$$

$\mathbf{x}_0$ is observable if and only if (Jazwinski 1970, p. 232)

$$\Phi = \mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k, 0) > 0. \tag{A.2}$$

Φ is called the information matrix with respect to $\mathbf{y}_k$.

If $\mathbf{L}(k, 0)$ is not invertible, then $\mathbf{x}_0$ must not be observable. However, even though $\mathbf{x}_0$ is not observable, the observation still reduces the analysis error variance since Φ is positive semi definite. Thus, lack of observability only implies that some sub-space is not observable. In the following, we will discuss properties of this unobservable sub-space.

Let us denote the rank of $\mathbf{L}(k, 0)$ as $q(k, 0)$. When $\mathbf{L}(k, 0)$ is not invertible, $q(k, 0)$ is less than its order n. In this case, $\mathbf{L}(k, 0)$ may be decomposed as

$$\mathbf{L}(k, 0) = \mathbf{U}(k, 0)\Sigma(k, 0)\mathbf{V}^{\mathrm{T}}(k, 0). \tag{A.3}$$

This is called a singular value decomposition. The $\mathbf{U}(k, 0)$, $\Sigma(k, 0)$, and $\mathbf{V}(k, 0)$ are determined as follows. We can have an orthonormal basis of eigenvectors from

$$\mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{L}(k, 0)\mathbf{v}_j = \sigma_j^2\mathbf{v}_j. \tag{A.4}$$

Note that $\mathbf{L}^{\mathrm{T}}(k, 0)\mathbf{L}(k, 0)$ is positive semi definite, and only has $q(k, 0)$ non-zero eigenvalues. $\mathbf{v}_j$ are also called singular vectors of $\mathbf{L}(k, i)$, and σ_j are the corresponding singular values. Then we introduce

$$\mathbf{u}_j = \frac{1}{\sigma_j}\mathbf{L}(k, i)\mathbf{v}_j, \quad j = 1, \ldots, q(k, i). \tag{A.5}$$

$\mathbf{u}_j$ also constitutes an orthonormal basis. By the Gram–Schmidt procedure, we complete $\mathbf{u}_j$ to a complete orthonormal basis of order n. Then, $\mathbf{V}(k, i)$ is an $n \times n$ matrix whose columns are the vectors $\mathbf{v}_j$, and $\mathbf{U}(k, 0)$ is also an $n \times n$ matrix whose columns are the vectors $\mathbf{u}_j$. Finally $\Sigma(k, 0)$ is an $n \times n$ diagonal matrix, and

$$\Sigma(k, 0) = \mathrm{diag}(\sigma_1, \ldots, \sigma_{q(k,0)}, 0, \ldots, 0). \tag{A.6}$$

We perform a space transformation in terms of the basis consisting of $\mathbf{v}_j$

$$\mathbf{z}_0 = \mathbf{V}^{\mathrm{T}}(k, 0)\mathbf{x}_0. \tag{A.7}$$

The corresponding Hessian becomes

$$\mathbf{H}_{0,l}^{\mathbf{z}_0} = \mathbf{V}^{\mathrm{T}}(k, 0)\mathbf{B}_0^{-1}\mathbf{V}(k, 0) + \Sigma(k, 0)\mathbf{U}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{U}(k, 0)\Sigma(k, 0), \tag{A.8}$$

and the information matrix becomes

$$\Phi = \Sigma(k, 0)\mathbf{U}^{\mathrm{T}}(k, 0)\mathbf{h}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{U}(k, 0)\Sigma(k, 0). \tag{A.9}$$

We see that the entries of Φ for $i, j > q(k, 0)$ are zero. Thus the sub space (rank space) spanned by the singular vectors associated with non-zero eigenvalues is observable, and the sub-space (null space) spanned by the singular vectors associated with zero eigenvalues is not observable.

It is noteworthy that an unobservable system is undetermined if there is no background term. On the contrary, the background term with a positive definite error covariance matrix is a sufficient condition for well posedness of the system. In this case, the background term is a regularization term as proposed by Tikhonov and Arsenin (1977).

REFERENCES

Anderson, B. D. O. and Moore, J. B. 1979 *Optimal filtering*. Prentice-Hall, Englewood Cliffs, USA

Bennett, A. F. 1992 *Inverse methods in physical oceanography*. Cambridge University Press, Cambridge, UK

Bennett, A. and Budgell, P. 1989 The Kalman smoother for a linear quasi-geostrophic model of ocean circulation model. *Dyn. Atmos. Oceans*, **13,** 219–267

Berre, L., Gustafsson, N. and Lindskog, M. 1999 'A comparison between an analytical and a statistical formulation of the background error balance in the HIRLAM 3D-Var'. In Proceedings of the Third WMO international symposium of observations in meteorology and oceanography, 7-11 June 1999, Québec City, Canada

Bryson, A. E. and Ho, Y.-C. 1975 *Applied optimal control*. Hemisphere Publishing Corporation, Washington DC, USA

Chao, W. C. and Chang, L.-P. 1992 Development of a four-dimensional variational analysis system using the adjoint method at GLA. Part 1: Dynamics. *Mon. Weather Rev.*, **120,** 1661–1674

Cohn, S. E. 1997 Estimation theory for data assimilation problems: Basic conceptual framework and some open questions. *J. Meteorol. Soc. Jpn.*, **75, 1B,** 257–288

1993 Dynamics of short-term univariate error covariances. *Mon. Weather Rev.*, **121,** 3123–3149

Cohn, S. E., da Silva, D. M., Guo, J., Sienkiewicz, M. and Lamich, D. 1998 Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Weather Rev.*, **126,** 2913–2926

Cohn, S. E. and Dee, D. 1988 Observability of discretized partial differential equations. *SIAM J. Numer. Anal.*, **25,** 586–617

Cohn, S. E. and Parrish, D. F. 1991 The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Weather Rev.*, **119,** 1757–1785

Cohn, S. E., Sivakumaran, N. S. and Todling, R. 1994 A fixed-lag Kalman smoother for retrospective data assimilation. *Mon. Weather Rev.*, **122,** 2838–2867

Courtier, P. 1997 Variational methods. *J. Meteorol. Soc. Jpn.*, **75, 1B,** 211–217

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F. and Fisher, M. 1998 The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.*, **124,** 1783–1807

Courtier, P. and Talagrand, O. 1987 Variational assimilation of meteorological observations with the adjoint vorticity equation. II. Numerical results. *Q. J. R. Meteorol. Soc.*, **113,** 1129–1347

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1994 A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120,** 1367–1388

Daley, R. 1991 *Atmospheric data assimilation*. Cambridge atmospheric and space science series, Cambridge University Press, Cambridge, USA

1992 The effect of serially correlated observations and model error on atmospheric data assimilation. *Mon. Weather Rev.*, **120,** 164–177

Dee, D. P. 1991 Simplification of the Kalman filter for meteorological data assimilation. *Q. J. R. Meteorol. Soc.*, **117,** 365–384

Dee, R. P. and da Silva, D. M. 1998 Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **124,** 269–298

Derber, J. C. 1989 A variational continuous assimilation technique. *Mon. Weather Rev.*, **117,** 2437–2446

Derber, J. C., Parrish, D. F. and Lord, S. J. 1991 The new global operational analysis system at the National Meteorological Center. *Weather Forecasting*, **6,** 538–547

Ehrendorfer, M. and Bouttier, F. 1998 'An explicit low-resolution extended Kalman filter: Implementation and preliminary experimentation'. Technical memorandum no. 259, ECMWF

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),** 10143–10162

Fisher, M. 1998 'Development of a simplified Kalman filter'. Technical memorandum no. 260, ECMWF

Fisher, M. and Courtier, P. 1995 'Estimating the covariance matrices analysis and forecast error in variational data assimilation'. Technical memorandum no. 222, ECMWF

Gauthier, P. C., Cherette, L., Fillion, L., Koclas, P. and Laroche, S. 1998 Implementation of a 3D variational data assimilation system at the Canadian Meteorological Center. Part I: The global analysis. *Atmos. Ocean*, **37,** 103–156

Ghil, M. 1989 Meteorological data assimilation for oceanographers. Part I. Description and theoretical framework. *Dyn. Atmos. Oceans*, **13,** 171–218

1997 Advances in sequential estimation for atmospheric and oceanic flows. *J. Meteorol. Soc. Jpn.*, **75, 1B,** 289–304

Ghil, M., Cohn, S. E., Tavantzis, J., Bube, K. and Isaacson, E. 1981 Application of estimation theory to numerical weather prediction. Pp. 139–224 in *Dynamical meteorology: Data assimilation methods*. Eds. L. Bengtsson, M. Ghil and E. Kallen, Spring-Verlag, New York, USA

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

Gibson, J. K., Hernandez, A., Kallberg, P., Nomura, A., Serrano, E. and Uppala, S. 1996 'Current status of the ECMWF re-analysis (ERA) project'. Pp. 283–290 in Proceedings of the Fifth workshop on meteorological operational systems, 13–17 November 1995, Reading, UK. ECMWF workshop proceedings, ECMWF, Reading, UK

Ingleby, N. B. S., Ballard, S., Clayton, A., Lorenc, A. C., Li, D. and Payne, T. 1999 'The effect of changes to the error covariances and observation weighting on a 3D-Var system'. In Proceedings of the Third WMO international symposium of observations in meteorology and oceanography, 7–11 June, Québec City, Canada

Jarvinnen, H., Thépaut, J.-N. and Courtier, P. 1996 Quasi-continuous variational data assimilation. *Q. J. R. Meteorol. Soc.*, **122,** 515–534

Jazwinski, A. H. 1970 *Stochastic processes and filtering theory*. Academic Press, New York, USA

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D: J. Basic Eng.*, **82,** 35–45

Kalnay, E., Kanamitsu, M., Kistler, K., Collin, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D. 1996 The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.*, **77,** 437–471

Le Dimet, F.-X. and Talagrand, O. 1986 Variational algorithms for analysis and assimilation of meteorological observations. *Tellus*, **38A,** 97–110

Lewis, J. and Derber, J. 1985 The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A,** 309–327

Li, Z., Navon, I. M. and Zhu, Y. 2000 Performance of 4D-Var with different strategies on the use of adjoint physics with the FSU global spectral model. *Mon. Weather Rev.*, **128,** 668–688

Lorenc, A. C. 1986 Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112,** 1177–1194

Mahfouf, J.-F. and Rabier, F. 2000 The ECMWF operational implementation of four-dimensional variational assimilation. Part II: Experimental results with improved physics. *Q. J. R. Meteorol. Soc.*, **126,** 1171–1190

Ménard, R. and Daley, R. 1996 The application of Kalman smoother theory to the estimation of 4DVAR error statistics. *Tellus*, **48A,** 221–237

Molteni, F. and Palmer, T. N. 1993 Predictability and finite-time instability of the northern winter circulation. *Q. J. R. Meteorol. Soc.*, **119,** 269–298

Navon, I. M., Zou, X., Derber, J. and Sela, J. 1992 Variational data assimilation with the NMC spectral model. Part 1: Adiabatic model tests. *Mon. Weather Rev.*, **120,** 1433–1446

Parrish, D. F. and Derber, J. C. 1992 The National Meteorological Center's spectral-interpolation system. *Mon. Weather Rev.*, **120,** 1747–1763

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

Rabier, F., Courtier, P., Pailleux, J., Talagrand, O. and Vasiljevic, D. 1993 A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on three-dimensional variational analysis. *Q. J. R. Meteorol. Soc.*, **119,** 845–880

Rabier, F., Järvinnen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A. 2000 The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126,** 1143–1170

Rabier, F., Thépaut, J.-N. and Courtier, P. 1998 Extended assimilation and forecast experiments with a four-dimensional variational assimilation system. *Q. J. R. Meteorol. Soc.*, **124,** 1861–1887

Schubert, S. D. and Rood, R. B. 1995 'Proceedings of the workshop on the GEOS-1 five-year assimilation'. NASA Tech. Memo. 104606, **Vol. 7**

Talagrand, O. and Courtier, P. 1987 Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Q. J. R. Meteorol. Soc.*, **113,** 1311–1328

Thépaut, J.-N., Alary, P., Caille, P., Cassé, V., Geleyn, J.-F., Moll, P., Pailleux, J., Piriou, J.-M., Puech, D. and Taillefer, F. 1998 'The operational global data assimilation system at Météo-France'. Pp. 25–31 in proceedings of HIRLAM4 workshop on variational analysis in limited area models, 1998, Toulouse, France

Thépaut, J.-N. and Courtier, P. 1991 Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Q. J. R. Meteorol. Soc.*, **117,** 1225–1254

Tikhonov, A. N. and Arsenin, V. 1977 *Solution of ill-posed problems.* Winston and Sons, Washington DC, USA

Todling, R. and Cohn, S. E. 1994 Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Mon. Weather Rev.*, **122,** 2530–2557

Todling, R., Cohn, S. E. and Sivakumaran, N. S. 1998 Suboptimal schemes for retrospective data assimilation based on the fixed-lag Kalman smoother. *Mon. Weather Rev.*, **126,** 2274–2286

Zhu, Y., Todling, R. and Cohn, S. E. 1999 'Technical remarks on smoother algorithms'. DAO Office Note 99-2, NASA/GSFC Data Assimilation Office

Zou, X. and Kuo, Y.-H. 1996 Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Mon. Weather Rev.*, **124,** 2859–2882

Zupanski, D. and Mesinger, F. 1995 Four-dimensional variational assimilation of precipitation data. *Mon. Weather Rev.*, **123,** 1112–1127

Zupanski, D. 1997 A general weak constraint applicable to operational 4DVAR data assimilation systems. *Mon. Weather Rev.*, **125,** 2274–2292

Zupanski, M. 1993 Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Weather Rev.*, **121,** 2396–2408