Optimality of 4D-Var and its relationship with the Kalman filter and Kalman smoother

Zhijin Li

Supercomputer Computations Research Institute

Florida State University, Tallahassee, Florida

I. M. Navon

Department of Mathematics and

Supercomputer Computations Research Institute

Florida State University, Tallahassee, Florida

submitted to **Q. J. Roy. Meteorol. Soc.**

June, 1999

Corresponding author:

I. Michael Navon

Supercomputer Computations Research Institute

Florida State University

Tallahassee, FL 32306-4130

Email: navon@scri.fsu.edu

Phone: (850) 644-6560

Fax: (850) 644-0098

Abstract

The known properties of equivalence between 4D-Var 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 method, the information in 4D-Var is fully transferred from one batch to the next by the background term; 4D-Var allows to process observations in subsets, while the final solution is optimal as all observations are processed simultaneously. These properties hold irrespective of whether the model is perfect or not. 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. Possible techniques for practical implementation of the equivalence are discussed.

Key words: Optimality, 4D-Var, Kalman filter, Kalman smoother

1 Introduction

The field of data assimilation for atmospheric sciences has witnessed an explosion of activity since the mid-80s. The desire to carry out data assimilation is being driven by the rapidly increasing amount of observational data becoming available, much of it from space-borne platforms, as well as by pressing scientific and societal needs to improve weather forecasting and understanding of new aspects of the behavior of the atmosphere. The most direct impetus is the development of physically and mathematically sound methodologies, along with advances in computing technology.

These methodologies comprise two families, that is, variational data assimilation and sequential data assimilation (Ghil and Malanotte-Rizzoli 1991, Ghil 1997; Daley 1991; Cohn 1997; Courtier 1997). The variational data assimilation includes 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. The framework of estimation theory constitutes the mathematical foundation for both variational and sequential data assimilation, though variational methods are closer to aspects of control theory (Cohn 1997; Lorenc 1986; Ghil and Malanotte-Rizzoli 1991; and references therein).

The variational data assimilation has undergone a rapid development since the papers 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 method (OI) in 1991 at the National Center for Environment Prediction (NCEP) (Parrish and Derber 1992; Derber et al. 1991), and then in 1996 at the European Center for Medium-Range Weather Forecasts (ECMWF) (Courtier et al. 1998), as well as in 1998 at the Data Assimilation Office (Cohn et al. 1998). In the same time, 4D-Var is extensively investigated and developed associated with operational models (e.g., Thepaut and Courtier 1991; Navon et al. 1992; Chao and Chang 1992; M. 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 has been implemented for operational use at the end of 1997.

Parallel to variational methods, sequential methods have attracted extensive attention since the beginning of 80's (Ghil et al. 1981). Most of meteorological applications of the Kalman filter use idealized one- and two-dimensional models (e.g., Cohn and Parrish 1991; Dee 1990; Todling and Cohn 1994; Daley 1992; Evensen 1994). Cohn et al. (1994) introduce a Kalman smoother to meteorology, called retrospective data assimilation, aiming to produce a time consistent long sequence of analyses. This method has also been studied in idealized one- and two-dimensional models (Cohn et al., 1994; Todling et al. 1998).

In estimation theory, variational methods are called the statistical method or least-squares approach (Jazwinski 1970, p151). The least squares approach does not seem to contribute much to estimation theory. However, it is fair to say that variational methods are currently dominating the scene of meteorological data assimilation. This phenomenon can primarily be attributed to two facts, namely, the large size of atmospheric models and the availability of advanced minimization algorithms which permit implementation of variational methods for operational purposes on available computer platforms. 4D-Var has been implemented using numerical weather prediction (NWP) models with the most sophisticated physical processes (e.g. D. Zupanski 1997; Zou 1997; Zou and Kuo 1996; Li and Navon 1999, Rabier et al. 1999; Mahfouf and Rabier 1999; Klinker et al. 1999).

The large size of meteorological models has prevented the Kalman filter and smoother from being implemented in their full form, since they require computation of time evolution of covariances matrices. While many advances have been made in developing suboptimal Kalman filters and smoothers, enormous efforts are still required to implement these suboptimal systems for practical use in meteorology (e.g., Todling et al. 1998; Bouttier 1996).

The connections between variational methods and sequential methods 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). The Kalman filter is optimal in the sense of producing the best linear unbiased estimate of the model state at the end of the 4D-var assimilation period (e.g., Jazwinski 1970). The equivalence between 4D-Var and the fixed-interval Kalman smoother has also been known even for models with error (Bryson and Ho 1975; Bennett and Budgell 1989; Menard and Daley 1996).

The equivalence between sequential and variational methods provides a possibility of combining these two methods. Such a combination may be expected to reduce the deficiencies in each method. It has been recognized that using simplified Kalman filters constitutes a promising approach for evaluating the propagation of covariance matrices (Menard and Daley 1996; Ehrendorfer and Bouttier 1998). This implies that using simplified Kalman filters may improve variational data assimilation. Especially, such methods have shown a prospect for practical use (Rabier et al., 1997; Fisher 1998).

As mentioned previously, the equivalence in many aspects between sequential and variational methods has been known. Recently, Cohn et al. (1998) have presented a comprehensive and consistent derivation of the equivalnce between these two methods. In this work, we focus on presenting a consistent and detailed analysis of the optimality of 4D-Var solutions, and then describe the equivalences between these two methods. We will prove that the 4D-Var solution is optimal not only with respect to the model trajectory consisting of the model states over the assimilation time interval, but also with respect to any model state at a single observation time irrespective of whether the model is perfect or not. We will further show that the information in 4D-Var can be fully transferred from one batch to the next by the background term when batch processing methods are used. Also, 4D-Var allows to process observations in subsets, while the final solution is optimal as all observations are processed simultaneously. These optimality properties of 4D-Var can be used to explain various properties of equivalence between 4D-Var and the Kalman filter/smoother. The results can be used as a baseline for evaluating approaches for combining the variational method and sequential method. Especially, it may constitute a starting point for constructing an approach equivalent to the fixed-lag Kalman smoother, aimed at producing a time-consistent long sequence of assimilated analyses based on the 4D-Var method, as substantiated in section 5.

The outline of this paper is as follows. In section 2, we briefly summarize 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 method and sequential method, and a 4D-Var smoother is proposed for producing a time-consistent long sequence of assimilated analyses. Finally, section 6 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,

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

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

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

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,

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

and $\mathbf{x}_0$ is independent of $\{\mathbf{w}_k\}$. We can see that the random sequence generated by (1) is Markov, since $\mathbf{x}_{k+1}$ is determined by $\mathbf{x}_k$ and $\mathbf{w}_k$, independent of $\{\mathbf{x}_{k-1}, \cdots, \mathbf{x}_0\}$ (Jazwinski 1971, p86). 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 $\{\mathbf{v}_k, k = 1, \cdots\}$ is an m-vector, white Gaussian sequence, $\mathbf{v}_k \sim N(0, \mathbf{R}_k)$, $\mathbf{R}_k > 0$. $\{\mathbf{w}_k\}$ 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

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

Given a realization of the sequence of observations $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$, that is, give $\mathbf{Y}_l$, 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 filtering problem; and if $k > l$, it is called the prediction problem.

It is clear that the conditional probability density function of $\mathbf{x}_k$ given $\mathbf{Y}_l$, which we write as

$$p(\mathbf{x}_k | \mathbf{Y}_l) \tag{6}$$

is the complete solution.

Let $\hat{\mathbf{x}}_k$ be an estimate of $\mathbf{x}_k$ given $\mathbf{Y}_l$. The error in the estimate is simply

$$\tilde{\mathbf{x}}_k \equiv \mathbf{x}_k - \hat{\mathbf{x}}_k \tag{7}$$

A good estimate should be clearly have the property of producing small errors.

Estimation theory shows that the minimum mean square error or minimum variance estimate is the conditional mean (Cohn 1997; Jazwinski 1970). It is noteworthy that this result is general, 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 the maximum likelihood (Bayesian) estimate obtained by maximizing $p(\mathbf{x}_k | \mathbf{Y}_l)$. When $p(\mathbf{x}_k | \mathbf{Y}_l)$ is Gaussian, the maximum likelihood (Bayesian) estimate is identical to the minimum variance estimate. In the followings, we shall use terminologies associated with the maximum likelihood (Bayesian) estimate to analyze both the variational and sequential methods (e.g., Jazwinski 1970).

3 Properties of 4D-Var

a Perfect model

Consistent property on 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 $\mathbf{x}^b$ and to the observations y distributed over a time interval $[t_0, t_l]$ (Lorenc 1986),

$$J^p = \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k). \quad (8)$$

We assume that there is no observation at the time t_0.

In the 4D-Var, the minimization of the cost function (8) is carried out with respect to the initial state $\mathbf{x}_0$. However, the minimization can be performed with respect to any $\mathbf{x}_k$. We show here that the choice of $\mathbf{x}_k$ does not make any difference in the solution.

The cost function (8) with respect to $\mathbf{x}_0$ is

$$J^p(\mathbf{x}_0) = \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{k=1}^{l}(\mathbf{h}_k\mathbf{L}(k,0)\mathbf{x}_0 - \mathbf{y}_k)^T \mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{L}(k,0)\mathbf{x}_0 - \mathbf{y}_k). \quad (9)$$

Here we have used the expression

$$\mathbf{L}(k,i) = \mathbf{L}(k,k-1)\cdots\mathbf{L}(i+1,i), \quad \text{for} \quad i \le k \quad (10)$$

The gradient of $J^p(\mathbf{x}_0)$ with respect to $\mathbf{x}_0$ can be written as

$$\nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0) = \mathbf{B}_0^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \sum_{k=1}^{l}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{L}(k,0)\mathbf{x}_0 - \mathbf{y}_k), \quad (11)$$

To find the minimum solution, let

$$\nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0) = 0 \tag{12}$$

Then we have the optimal solution

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0^b - \mathbf{H}_{0,l}^{-1} \nabla_{\mathbf{x}_0} J(\mathbf{x}_0^b). \tag{13}$$

Here

$$\mathbf{H}_{0,l} = \mathbf{B}_0^{-1} + \sum_{k=1}^{l} \mathbf{L}^T(k,0)\mathbf{h}_k^T R_k^{-1} \mathbf{h}_k \mathbf{L}(k,0) \tag{14}$$

is the Hessian matrix of the cost function (9) at $\mathbf{x}_0$, and

$$\nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0^b) = \sum_{k=1}^{l} \mathbf{L}^T(k,0)\mathbf{h}_k^T \mathbf{R}_k^{-1}(\mathbf{y}_k^b - \mathbf{y}_k), \tag{15}$$

the gradient of the cost function (9) at $\mathbf{x}_0^b$, where

$$\mathbf{y}_k^b = \mathbf{h}_k \mathbf{x}_k^b = \mathbf{h}_k \mathbf{L}(k,0)\mathbf{x}_0^b. \tag{16}$$

The Hessian is a semi-positive definite symmetric matrix. We note also that $-\mathbf{H}_{0,l}^{-1} \nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0^b)$ is the Newton direction in optimal control theory.

Following Rabier et al. (1992), we can obtain the analysis error covariance matrix

$$P_0^a = < \hat{\mathbf{x}}_0 - \mathbf{x}_0^t >< \hat{\mathbf{x}}_0 - \mathbf{x}_0^t >^T = \mathbf{H}_{0,l}^{-1}, \tag{17}$$

where $\mathbf{x}_0^t$ is the assumed true value of $\mathbf{x}_0$. It is interesting that the error covariance is determined by the second derivative of the cost function.

We turn to the cost function (8) with respect to $\mathbf{x}_k$, $(k = 1, \cdots, l)$, which can be written as

$$\begin{aligned} J^p(\mathbf{x}_k) \quad = \quad & \frac{1}{2}[\mathbf{x}_k - \mathbf{x}_k^b]^T \mathbf{L}^{-T}(k,0)\mathbf{B}_0^{-1}\mathbf{L}(k,0)[\mathbf{x}_k - \mathbf{x}_k^b] \\ & + \frac{1}{2}\sum_{i=1}^{k}[\mathbf{h}_i \mathbf{L}^{-1}(k,i)\mathbf{x}_k - \mathbf{y}_i]^T \mathbf{R}_i^{-1}[\mathbf{h}_i \mathbf{L}^{-1}(k,i)\mathbf{x}_k - \mathbf{y}_i] \\ & + \frac{1}{2}\sum_{i=k+1}^{l}[\mathbf{h}_i \mathbf{L}(i,k)\mathbf{x}_k - \mathbf{y}_i]^T \mathbf{R}_i^{-1}[h_i \mathbf{L}(i,k)\mathbf{x}_k - \mathbf{y}_i]. \end{aligned} \tag{18}$$

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

Its gradient with respect to $\mathbf{x}_k$ can be written as

$$\begin{aligned}
\nabla_{\mathbf{x}_k} J^p(\mathbf{x}_k) &= \mathbf{L}^{-T}(k,0)\mathbf{B}_0^{-1}\mathbf{L}(k,0)[\mathbf{x}_k - \mathbf{x}_k^b] \\
&\quad + \sum_{i=1}^{k} \mathbf{L}^{-T}(k,i)\mathbf{h}_i^T \mathbf{R}_i^{-1}[\mathbf{h}_i \mathbf{L}^{-1}(k,i)\mathbf{x}_k - \mathbf{y}_i] \\
&\quad + \sum_{i=k+1}^{l} \mathbf{L}^T(i,k)\mathbf{h}_i^T \mathbf{R}_i^{-1}[\mathbf{h}_i \mathbf{L}(i,k)\mathbf{x}_k - \mathbf{y}_i]
\end{aligned} \tag{19}$$

A process similar to that for $\mathbf{x}_0$ yields the optimal solution

$$\hat{\mathbf{x}}_k = \mathbf{x}_k^b + \mathbf{H}_{k,l}^{-1} \nabla_{\mathbf{x}_k} J(\mathbf{x}_k^b), \tag{20}$$

Here

$$\begin{aligned}
\mathbf{H}_{k,l} &= \mathbf{L}^{-T}(k,0)\mathbf{B}_0^{-1}\mathbf{L}^{-1}(k,0) + \sum_{i=1}^{k} \mathbf{L}^{-T}(k,i)\mathbf{h}_i^T \mathbf{R}_i^{-1}\mathbf{h}_i \mathbf{L}^{-1}(k,i) \\
&\quad + \sum_{i=k+1}^{l} \mathbf{L}^T(i,k)\mathbf{h}_i^T \mathbf{R}_i^{-1}\mathbf{h}_i \mathbf{L}(i,k)
\end{aligned} \tag{21}$$

is the Hessian matrix of the cost function (18) at $\mathbf{x}_k$, and

$$\begin{aligned}
\nabla_{\mathbf{x}_k} J^p(\mathbf{x}_k^b) &= \sum_{i=1}^{k} \mathbf{L}^{-T}(k,i)\mathbf{h}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{y}_i^b) \\
&\quad + \sum_{i=k+1}^{l} \mathbf{L}(i,k)^T \mathbf{h}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{y}_i^b)
\end{aligned} \tag{22}$$

is the gradient of the cost function (18) at $\mathbf{x}_k^b$.

We can obtain the analysis error covariance matrix

$$\mathbf{P}_k^a = < \hat{\mathbf{x}}_k - \mathbf{x}_k^t >< \hat{\mathbf{x}}_k - \mathbf{x}_k^t >^T = \mathbf{H}_{k,l}^{-1}, \tag{23}$$

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

We further obtain

$$\mathbf{L}^T(k,0)\mathbf{H}_{k,l}\mathbf{L}(k,0) = \mathbf{H}_{0,l} \tag{24}$$

$$\mathbf{L}^T(k,0)\nabla_{\mathbf{x}_k} J^p(\mathbf{x}_k^b) = \nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0^b) \tag{25}$$

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

$$\mathbf{P}_k^a = \mathbf{L}(k,0)\mathbf{P}_0^a\mathbf{L}^T(k,0) \tag{27}$$

The results expressed by (26) and (27) show that the optimal solution and the associated analysis error covariance are propagated by the linear model.

Additive property

For the cost function (8), we process all the observations simultaneously. We may separate the observations into a few subsets. The 4D-Var process is performed for each subset. The final solution can be the same as the one when all the observation is simultaneously processed.

We define two cost functions

$$J_1^p(\mathbf{x}_0) = \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T\mathbf{B}_0^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{i=1}^{l_1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_i} - \mathbf{y}_{k_1(i)})^T\mathbf{R}_{k_1(i)}^{-1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_i} - \mathbf{y}_{k_1(i)}), \tag{28}$$

$$J_2^p(\mathbf{x}_0) = \frac{1}{2}[\mathbf{x}_0 - \hat{\mathbf{x}}_0^b]^T\hat{\mathbf{B}}_0^{-1}[\mathbf{x}_0 - \hat{\mathbf{x}}_0^b] + \frac{1}{2}\sum_{i=1}^{l_2}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)})^T\mathbf{R}_{k_2(i)}^{-1}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)}), \tag{29}$$

where $l = l_1 + l_2$, and $k_1(i)$ $(i = 1, \cdots, l_1)$ and $k_2(i)$ $(i = 1, \cdots, l_2)$ make up of k $(k = 1, \cdots, l)$. The 4D-Var calculation is first carried out with J_1^p with respect to $\mathbf{x}_0$. The optimal solution $\hat{\mathbf{x}}_0^b$ is used as the background in J_2^p. $\hat{\mathbf{B}}_0^{-1}$ is the analysis error covariance associated with $\hat{\mathbf{x}}_0^b$. Then the 4D-Var calculation is carried out with J_2^p. The final solution $\hat{\hat{\mathbf{x}}}_0^b$ is the same as the one that directly minimizes (8).

We can prove this property by directly examining the minimization solution. Here we present a concise proof. Using the procedure similar to the above, we obtain the minimization solution of J_1^p

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0^b - \mathbf{H}_{0,l_1}^{-1}\nabla_{\mathbf{x}_0} J_1^p(\mathbf{x}_0^b), \tag{30}$$

where

$$\mathbf{H}_{0,l_1} = \mathbf{B}_0^{-1} + \sum_{i=1}^{l_1} \mathbf{L}^T(k_1(i),0)\mathbf{h}_{k_1(i)}^T R_{k_1(i)}^{-1} \mathbf{h}_{k_1(i)} \mathbf{L}(k_1(i),0) \tag{31}$$

is the Hessian matrix of the cost function (28) at $\mathbf{x}_0$, and

$$\nabla_{\mathbf{x}_0} J^p(\mathbf{x}_0^b) = \sum_{i=1}^{l_1} \mathbf{L}^T(k_1(i),0)\mathbf{h}_{k_1(i)}^T \mathbf{R}_{k_1(i)}^{-1} (\mathbf{y}_{k_1(i)}^b - \mathbf{y}_{k_1(i)}) \tag{32}$$

the gradient of the cost function (28) at $\mathbf{x}_0^b$.

Thus, J_1^p may be written as

$$J_1^p(\mathbf{x}_0) = (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T \mathbf{H}_{0,l_1} (\mathbf{x}_0 - \hat{\mathbf{x}}_0) + J_1^p(\hat{\mathbf{x}}_0) \tag{33}$$

Viewing $\mathbf{H}_{0,l_1}$ in (33) as $\hat{\mathbf{B}}_0^{-1}$ in (29), we obtain

$$J^p(\mathbf{x}_0) = J_2^p(\mathbf{x}_0) + J_1^p(\hat{\mathbf{x}}_0). \tag{34}$$

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 $J_2^p(\mathbf{x}_0)$.

b Model with errors

Consistent property on the time interval

When the model error are considered, the cost function in 4D-Var can be written as (Jazwinski 1970; Cohn 1997)

$$\begin{aligned} J^e &= \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1} [\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{k=1}^{l} (\mathbf{h}_k \mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{x}_k - \mathbf{y}_k) \\ &\quad + \frac{1}{2}\sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k. \end{aligned} \tag{35}$$

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 D. Zupanski 1997).

We have three choices for carrying out minimization of the cost function (35), that is,

1. Carrying out constrained minimization J^e with respect to $(\mathbf{x}_0, \cdots, \mathbf{x}_l, \mathbf{w}_1, \cdots, \mathbf{w}_l)$, subject to (1).

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

$$\begin{aligned} J_1^e &= \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1} [\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{k=1}^{l} (\mathbf{h}_k \mathbf{x}_k - \mathbf{y}_k)^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{x}_k - \mathbf{y}_k) \\ &+ \frac{1}{2}\sum_{k=1}^{l} (\mathbf{x}_k - \mathbf{L}(k, k-1)\mathbf{x}_{k-1})^T \mathbf{Q}_k^{-1} (\mathbf{x}_k - \mathbf{L}(k, k-1)\mathbf{x}_{k-1}). \end{aligned} \tag{36}$$

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

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

$$\begin{aligned} J_2^e &= \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1} [\mathbf{x}_0 - \mathbf{x}_0^b] \\ &+ \frac{1}{2}\sum_{k=1}^{l} \{\mathbf{h}_k [\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j] - \mathbf{y}_k\}^T \\ &\quad \mathbf{R}_k^{-1} \{\mathbf{h}_k [\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j] - \mathbf{y}_k\} \\ &+ \frac{1}{2}\sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k. \end{aligned} \tag{37}$$

   with respect to $\{\mathbf{x}_0, \mathbf{w}_1 \cdots, \mathbf{w}_l\}$.

In the derivation of J_2^e, we have used the following relationship

$$\mathbf{x}_k = \mathbf{L}(k,0)x_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j. \tag{38}$$

We start our discussion from J_2^e.

The gradient of J_2^e with respect to $\{\mathbf{x}_0, \mathbf{w}_1 \cdots, \mathbf{w}_l\}$ is given by

$$\begin{aligned}
\nabla_{\mathbf{x}_0} J_2^e &= \mathbf{B}^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] \\
&+ \sum_{k=1}^{l} \mathbf{L}^T(k,0)\mathbf{h}_k^T \mathbf{R}_k^{-1}\{\mathbf{y}_k - \mathbf{h}_k[\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j]\} \qquad (39) \\
\nabla_{\mathbf{w}_k} J_2^e &= \mathbf{Q}_k^{-1}\mathbf{w}_k + \sum_{j=k}^{l}\Gamma^T(j-1)\mathbf{L}^T(l,j)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k), \qquad (40)
\end{aligned}$$

where $k = 1, \cdots, l$.

Let

$$\nabla_{\mathbf{x}_0} J_2^e = 0. \qquad (41)$$

We obtain

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0^b + \mathbf{H}_{0,l}^{-1}\nabla_{\mathbf{w}_k} J_2^e(\mathbf{x}_0^b). \qquad (42)$$

Here $H_{0,l}$ is defined by (14), and

$$\nabla_{\mathbf{w}_k} J_2^e(\mathbf{x}_0^b) = \sum_{k=0}^{l}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{y}_k^b) \qquad (43)$$

where

$$\mathbf{y}_k^b = \mathbf{h}_k[\mathbf{L}(k,0)\mathbf{x}_0^b + \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\hat{\mathbf{w}}_j] \qquad (44)$$

Similar to deriving (27) for the perfect model, we can obtain the analysis error covariance

$$\begin{aligned}
\mathbf{P}_0^a &= \mathbf{H}_{0,1}^{-1} + \mathbf{H}_{0,1}^{-1}\{\sum_{k=1}^{l}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k \\
&\quad [\sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\hat{\mathbf{Q}}_j\Gamma^T(j-1)\mathbf{L}^T(k,j)]\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k,0)\}\mathbf{H}_{0,1}^{-1}, \qquad (45)
\end{aligned}$$

where $\hat{\mathbf{Q}}_j$ is the covariance of $\hat{\mathbf{w}}_j$. We note that the second term on the right hand is positive semi-definite, induced by the model error. This term reflects the fact that model errors always lead to an increase in the analysis error.

As has been noted for the perfect model, we can minimize J_2^e with respect to $(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l)$, rather than $(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l)$. The similar process as done for the perfect model yields

$$\hat{\mathbf{x}}_k = \mathbf{L}(k,0)\hat{\mathbf{x}}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\hat{\mathbf{w}}_j \tag{46}$$

where $\hat{\mathbf{w}}_j$ is the optimal value of $\mathbf{w}_j$, independent of the choice of $\mathbf{x}_k$.

We have seen that it is hard to obtain the analytic minimum solution of J_2^e, but the following expressions of the minimum solution can be useful for describing its properties.

Let $\mathbf{Z}^T = (\mathbf{x}_0^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_l^T)$. Then J_2^e can be written as

$$J_2^e(\mathbf{Z}) = \frac{1}{2}\mathbf{Z}^T H_{\mathbf{Z}}\mathbf{Z} + \mathbf{r}\mathbf{Z} + \delta, \tag{47}$$

where $\mathbf{H}_{\mathbf{Z}}$, $\mathbf{r}$ and δ are determined respectively by

$$\begin{aligned}
\mathbf{Z}^T\mathbf{H}_{\mathbf{Z}}\mathbf{Z} &= \mathbf{x}_0^T\mathbf{B}_0^{-1}\mathbf{x}_0 \\
&+ \sum_{k=1}^{l}[\mathbf{h}_k[\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j]^T \\
&\quad \mathbf{R}_k^{-1}[\mathbf{h}_k\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j] \\
&+ \sum_{k=1}^{l}\mathbf{w}_k^T\mathbf{Q}_k^{-1}\mathbf{w}_k,
\end{aligned} \tag{48}$$

$$\begin{aligned}
\mathbf{r}\mathbf{Z} &= -\{(\mathbf{x}_0^b)^T\mathbf{B}_0^{-1}\mathbf{x}_0 \\
&+ \sum_{k=1}^{l}(\mathbf{y}_k)^T\mathbf{R}_k^{-1}[\mathbf{h}_k[\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j]\},
\end{aligned} \tag{49}$$

$$\delta = \frac{1}{2}[(\mathbf{x}_0^b)^T\mathbf{B}_0^{-1}\mathbf{x}_0^b + (\mathbf{y}_k)^T\mathbf{R}_k^{-1}\mathbf{y}_k]. \tag{50}$$

$\mathbf{H}_{\mathbf{Z}}$ is the Hessian of $J_2^e(\mathbf{Z})$ with respect to $\mathbf{Z}$. From (48), it follows that $\mathbf{H}_{\mathbf{Z}}$ is symmetric positive-definite. Thus, $J_2^e(\mathbf{Z})$ possesses an unique minimum solution, that is,

$$\hat{\mathbf{Z}} = -\mathbf{H}_{\mathbf{Z}}^{-1}\mathbf{r}. \tag{51}$$

Then we have

$$J_2^e(\mathbf{Z}) = \frac{1}{2}(\mathbf{Z} - \hat{\mathbf{Z}})^T \mathbf{H}_{\mathbf{Z}} (\mathbf{Z} - \hat{\mathbf{Z}}) + J_2^e(\hat{\mathbf{Z}}) \tag{52}$$

Additive property

4D-Var with models where errors are present also holds the additive property similar to that for perfect models. We define two cost functions

$$\begin{aligned} J_{2,1}^e(\mathbf{Z}) &= \frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1}[\mathbf{x}_0 - \mathbf{x}_0^b] + \frac{1}{2}\sum_{i=1}^{l_1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_i} - \mathbf{y}_{k_1(i)})^T \mathbf{R}_{k_1(i)}^{-1}(\mathbf{h}_{k_1(i)}\mathbf{x}_{k_i} - \mathbf{y}_{k_1(i)}) \\ &\quad + \sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k, \end{aligned} \tag{53}$$

$$J_{2,2}^e(\mathbf{Z}) = \frac{1}{2}[\mathbf{Z} - \hat{\mathbf{Z}}_1]^T \hat{\mathbf{B}}_{\mathbf{Z}}^{-1}[\mathbf{Z} - \hat{\mathbf{Z}}_1] + \frac{1}{2}\sum_{i=1}^{l_2}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)})^T \mathbf{R}_{k_2(i)}^{-1}(\mathbf{h}_{k_2(i)}\mathbf{x}_{k_2(i)} - \mathbf{y}_{k_2(i)}). \tag{54}$$

When $\hat{\mathbf{Z}}_1$ is the minimization solution $J_{2,1}^e(\mathbf{Z})$ and $\hat{\mathbf{B}}_{\mathbf{Z}_1}^{-1}$ is taken as $\mathbf{H}_{\mathbf{Z}_1}$, it follows from (52)

$$J_2^e(\mathbf{Z}) = J_{2,2}^e(\mathbf{Z}) + J_{2,1}^e(\hat{\mathbf{Z}}_1) \tag{55}$$

Thus, the minimization solution of $J_{2,2}^e(\mathbf{Z})$ is identical to that of $J_2^e(\mathbf{Z})$, where all the observations are processed simultaneously. This additive property of 4D-Var allows us to use observations subset by subset.

4 Optimality of 4D-Var

a Perfect model

For a perfect model, it has been shown that the minimum solution of the cost function (8) is the joint maximum likelihood (Bayesian) estimate, that is, the estimate is the mode of the joint conditional density. As such, it is called the maximum likelihood or most probable estimate. We provide here a concise derivation of this result following Lorenc (1986) and Jazwinski (1970, p153-155).

The maximum likelihood estimation consists in maximizing the conditional probability density function

$$p(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) \tag{56}$$

with respect to $(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l)$.

Using Bayes's rule, we write the density in (56) as

$$p(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) = \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \cdots, \mathbf{x}_l) p(\mathbf{x}_0, \cdots, \mathbf{x}_l)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)} \tag{57}$$

For a perfect model, $\mathbf{x}_k$, $(k = 1, \cdots, l)$ are determined by $\mathbf{x}_0$. Thus, we have

$$p(\mathbf{x}_0, \cdots, \mathbf{x}_l) = p(\mathbf{x}_0) \tag{58}$$

When $\mathbf{x}_0, \cdots, \mathbf{x}_l$ are given, $\mathbf{y}_1, \cdots, \mathbf{y}_l$ are independent due to (4). We get

$$p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \cdots, \mathbf{x}_l) = p(\mathbf{y}_1 | \mathbf{x}_1) \cdots p(\mathbf{y}_l | \mathbf{x}_l) = p_{\mathbf{v}_1}(\mathbf{y}_1 - \mathbf{h}_1 \mathbf{x}_1) \cdots p_{\mathbf{v}_l}(\mathbf{y}_l - \mathbf{h}_l \mathbf{x}_l), \tag{59}$$

where $p_{\mathbf{v}_i}$ are the probability density function with respect to $\mathbf{v}_i$.

We finally have

$$p(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) = c p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \cdots, \mathbf{x}_l) p(\mathbf{x}_0, \cdots, \mathbf{x}_l), \tag{60}$$

where c is only a function of Y_l, independent of $\{\mathbf{x}_0, \cdots, \mathbf{x}_l\}$.

From the assumptions described in section 2, (60) becomes

$$\begin{aligned} & p(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) = \\ & \quad c' \exp \left\{ \frac{1}{2} [\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1} [\mathbf{x}_0 - \mathbf{x}_0^b] \right. \\ & \quad \left. + \frac{1}{2} \sum_{k=1}^{n} (\mathbf{y}_k - \mathbf{h}_k \mathbf{x}_k)^T \mathbf{R}_k^{-1} (\mathbf{y}_k - \mathbf{h}_k \mathbf{x}_k) \right\}, \end{aligned} \tag{61}$$

where c' only depends on c. Thus, minimizing $p(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l)$ with respect to $\{\mathbf{x}_0, \cdots, \mathbf{x}_l\}$ is equivalent to minimizing (8).

The above solution is jointly optimal with respect to $\{x_0, \cdots, x_l\}$, that is, the model trajectory. Now we are concerned with the optimal solution with respect to $\mathbf{x}_k$ $(0 \leq k \leq l)$. It is obtained by maximizing the marginal density

$$p(\mathbf{x}_k | \mathbf{y}_1, \cdots, \mathbf{y}_l) \tag{62}$$

rather than the joint density function.

Using Bayes' rule, we get

$$p(\mathbf{x}_k | \mathbf{y}_1, \cdots, \mathbf{y}_l) = \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_k) p(\mathbf{x}_k)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)}. \tag{63}$$

Since x_0 is Gaussian,

$$\mathbf{x}_k = \mathbf{L}(k, 0)\mathbf{x}_0, \tag{64}$$

is also Gaussian, and its distribution is

$$\mathbf{x}_k \sim N\left(\mathbf{L}(k, 0)\mathbf{x}_0^b, \mathbf{L}(k, 0)\mathbf{B}_0\mathbf{L}^T(k, 0)\right). \tag{65}$$

In view of (4), we get

$$\mathbf{y}_i - \mathbf{h}_i\mathbf{L}^{-1}(k, i)\mathbf{x}_k = \mathbf{v}_i \quad i \leq k \tag{66}$$

$$\mathbf{y}_i - \mathbf{h}_i\mathbf{L}(i, k)\mathbf{x}_k = \mathbf{v}_i \quad i > k \tag{67}$$

Thus, with $\{\mathbf{x}_0, \cdots, \mathbf{x}_l\}$ given, $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$ are independent. We have

$$\begin{aligned} & p(\mathbf{x}_k | \mathbf{y}_1, \cdots, \mathbf{y}_l) = \\ & \quad c p_{\mathbf{v}_1}(\mathbf{y}_1 - \mathbf{h}_1\mathbf{L}^{-1}(k, 1)\mathbf{x}_k) \cdots p_{\mathbf{v}_k}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}^{-1}(k, k)\mathbf{x}_k) \\ & \quad p_{\mathbf{v}_{k+1}}(\mathbf{y}_{k+1} - \mathbf{h}_{k+1}\mathbf{L}^{-1}(k+1, k)\mathbf{x}_k) \cdots p_{\mathbf{v}_l}(\mathbf{y}_l - \mathbf{h}_l\mathbf{L}^{-1}(l, k)\mathbf{x}_k) \end{aligned} \tag{68}$$

Then (68) becomes

$$
\begin{aligned}
&p(\mathbf{x}_k|\mathbf{y}_1,\cdots,\mathbf{y}_l) = \\
&\quad c' \exp\left\{-\frac{1}{2}[\mathbf{x}_k-\mathbf{x}_k^b]^T L^{-T}(k,0)\mathbf{B}_0^{-1}\mathbf{L}(k,0)[\mathbf{x}_k-\mathbf{x}_k^b]\right. \\
&\quad -\frac{1}{2}\sum_{i=0}^{k}[\mathbf{h}_i\mathbf{L}^{-1}(k,i)\mathbf{x}_k-\mathbf{y}_i]^T\mathbf{R}_i^{-1}[\mathbf{h}_i\mathbf{L}^{-1}(k,i)\mathbf{x}_k-\mathbf{y}_i] \\
&\quad \left.-\frac{1}{2}\sum_{i=k+1}^{l}[\mathbf{h}_i\mathbf{L}(i,k)\mathbf{x}_k-\mathbf{y}_i]^T\mathbf{R}_i^{-1}[\mathbf{h}_i\mathbf{L}(i,k)\mathbf{x}_k-\mathbf{y}_i]\right\}.
\end{aligned}
\tag{69}
$$

Maximizing $p(\mathbf{x}_k|\mathbf{y}_1,\cdots,\mathbf{y}_l)$ turns out to be equivalent to minimizing (18).

In view of the property of the 4D-Var solution described in section 3, we come to the conclusion that the 4D-Var solution is optimal not only with respect to the model trajectory consisting of $\{\mathbf{x}_0,\cdots,\mathbf{x}_l\}$ but also with respect to $\mathbf{x}_k$ at a single observation time.

b Model with errors

For a model with errors, the estimation problem constitutes finding the optimal model state and model forcing (control term) with the information from observations. From (38), we know that an observation $\mathbf{y}_k$ depends on $\{\mathbf{w}_1,\cdots,\mathbf{w}_k\}$, but is independent of $\{\mathbf{w}_{k+1},\cdots,\mathbf{w}_l\}$. This allows to improve the prior estimates of $\{\mathbf{w}_1,\cdots,\mathbf{w}_k\}$ in terms of the observations $\mathbf{Y}_k$.

Here we first consider to maximize

$$
p(\mathbf{x}_0,\mathbf{w}_1,\cdots,\mathbf{w}_l|\mathbf{y}_1,\cdots,\mathbf{y}_l). \tag{70}
$$

with respect to $\{\mathbf{x}_0,\mathbf{w}_1,\cdots,\mathbf{w}_l\}$. Similar to Jazwinski (1970, p153-155), we shall see that the 4D-Var solution minimizing the cost function (37) is equivalent to this maximum likelihood estimate.

Using Bayes' rule, we write the density in (70) as

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

$$= \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)}$$

$$= \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{x}_0 | \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{w}_1, \cdots, \mathbf{w}_l)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)} \tag{71}$$

With $\{\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l\}$ given, $\mathbf{x}_k$ is then determined by (38). In view of (4), with $\mathbf{x}_k$ given, $\mathbf{y}_k$ are Gaussian and independents, since $\mathbf{v}_k$ are white. We have

$$\begin{aligned} & p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l) \\ & \quad = p(\mathbf{y}_1 | \mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l) \cdots p(\mathbf{y}_l | \mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l) \\ & \quad = p(\mathbf{y}_1 | \mathbf{x}_1) \cdots p(\mathbf{y}_l | \mathbf{x}_l) \\ & \quad = p_{\mathbf{v}_1}[\mathbf{y}_1 - \mathbf{h}_1(\mathbf{L}(k,0)x_0 + \sum_{j=1}^{1} \mathbf{L}(1,j)\Gamma(j-1)\mathbf{w}_j)] \\ & \qquad \cdots p_{\mathbf{v}_l}[\mathbf{y}_l - \mathbf{h}_l(\mathbf{L}(k,0)x_0 + \sum_{j=1}^{l} \mathbf{L}(l,j)\Gamma(j-1)\mathbf{w}_j)] \end{aligned} \tag{72}$$

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

$$p(\mathbf{x}_0 | \mathbf{w}_1, \cdots, \mathbf{w}_l) = p_{\mathbf{x}_0}(\mathbf{x}_0) \tag{73}$$

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

$$p(\mathbf{w}_1, \cdots, \mathbf{w}_l) = p_{\mathbf{w}_1}(\mathbf{w}_1) \cdots p_{\mathbf{w}_l}(\mathbf{w}_l) \tag{74}$$

Substitution of (72), (73) and (74) into (71) yields

$$\begin{aligned} & p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) \\ & \quad = c' \exp\{ -\frac{1}{2}[\mathbf{x}_0 - \mathbf{x}_0^b]^T \mathbf{B}_0^{-1} [\mathbf{x}_0 - \mathbf{x}_0^b] \\ & - \quad \frac{1}{2} \sum_{k=1}^{l} \{[\mathbf{h}_k[\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j] - \mathbf{y}_k\}^T \\ & \quad \mathbf{R}_k^{-1} \{[\mathbf{h}_k \mathbf{L}(k,0)\mathbf{x}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j] - \mathbf{y}_k\} \\ & - \quad \frac{1}{2} \sum_{k=1}^{l} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k. \end{aligned} \tag{75}$$

where c' only depends on $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$, and is independent of $\{\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l\}$. We observe that maximizing $p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l)$ is equivalent to minimizing the cost function (37).

From (52), we can conclude that $p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l | Y_l)$ is Gaussian, and thus $p(\mathbf{x}_0|Y_l)$, $p(\mathbf{w}_1|Y_l)$, $\cdots$ $p(\mathbf{w}_l|Y_l)$ are also Gaussian. The minimum solution of (37), $\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_l$, provides the maximum likelihood estimates, that is, $\hat{\mathbf{x}}_0, \hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_l$ maximize $p(\mathbf{x}_0|Y_l)$, $p(\mathbf{w}_1|Y_l)$, $\cdots$ $p(\mathbf{w}_l|Y_l)$, respectively.

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

$$\begin{aligned} & p(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) \\ & \quad = \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)} \\ & \quad = \frac{p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{x}_k | \mathbf{w}_1, \cdots, \mathbf{w}_l) p(\mathbf{w}_1, \cdots, \mathbf{w}_l)}{p(\mathbf{y}_1, \cdots, \mathbf{y}_l)} \end{aligned} \tag{76}$$

From (1), we get

$$\mathbf{x}_k = \mathbf{L}(k, j)\mathbf{x}_j + \sum_{i=j}^{k} \mathbf{L}(k, i)\Gamma(i-1)\mathbf{w}_i \tag{77}$$

Thus , for $j \leq k$, we have

$$\mathbf{x}_j = \mathbf{L}^{-1}(k, j) \left(\mathbf{x}_k - \sum_{i=j}^{k} \mathbf{L}(k, i)\Gamma(i-1)\mathbf{w}_i \right). \tag{78}$$

Then

$$\begin{aligned} \mathbf{y}_j - \mathbf{h}_j\mathbf{x}_j &= \mathbf{y}_j - \mathbf{h}_j\mathbf{L}^{-1}(k, j) \left[\mathbf{x}_k - \sum_{i=j}^{k} \mathbf{L}(k, i)\Gamma(i-1)\mathbf{w}_i \right] \\ &= \mathbf{v}_j. \end{aligned} \tag{79}$$

For $j > k$, we have

$$\mathbf{x}_j = \mathbf{L}(j, k)x_k + \sum_{i=k}^{j} \mathbf{L}(i, k)\Gamma(i-1)\mathbf{w}_i. \tag{80}$$

Then

$$\begin{aligned}
\mathbf{y}_j - \mathbf{h}_j x_j &= \mathbf{y}_j - \mathbf{h}_j \left[\mathbf{L}(j,k) x_k + \sum_{i=k}^{j} \mathbf{L}(i,k)\Gamma(i-1)\mathbf{w}_i \right] \\
&= \mathbf{v}_j. \qquad (81)
\end{aligned}$$

In view of (79) and (81), with $\mathbf{x}_k$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_l\}$ given, $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$ are Gaussian and independent. Thus, we have

$$\begin{aligned}
& p(\mathbf{y}_1, \cdots, \mathbf{y}_l | \mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l) = \\
& \quad p_{\mathbf{v}_1}\left(\mathbf{y}_1 - \mathbf{h}_1 \mathbf{L}^{-1}(k,1) \left[\mathbf{x}_k - \sum_{i=1}^{k} \mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i \right] \right) \\
& \quad \cdots p_{\mathbf{v}_k}\left(\mathbf{y}_k - \mathbf{h}_k \mathbf{L}^{-1}(k,k) \left[\mathbf{x}_k - \sum_{i=k}^{k} \mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i \right] \right) \\
& \quad p_{\mathbf{v}_{k+1}}\left(\mathbf{y}_{k+1} - \mathbf{h}_{k+1} \left[\mathbf{L}(k+1,k)\mathbf{x}_k + \sum_{i=k}^{k+1} \mathbf{L}(i,k)\Gamma(i-1)\mathbf{w}_i \right] \right) \\
& \quad \cdots p_{\mathbf{v}_l}\left(\mathbf{y}_l - \mathbf{h}_l \left[\mathbf{L}(l,k)\mathbf{x}_k + \sum_{i=k}^{l} \mathbf{L}(i,k)\Gamma(i-1)\mathbf{w}_i \right] \right) \qquad (82)
\end{aligned}$$

In view of (38), with $\{\mathbf{w}_1, \cdots, \mathbf{w}_l\}$ given and $\mathbf{x}_0$ the Gaussian distribution, $\mathbf{x}_k$ is Gaussian. We obtain the probability density function

$$\begin{aligned}
p(\mathbf{x}_k | \mathbf{w}_1, \cdots, \mathbf{w}_l) &= p\left(\mathbf{L}(k,0)\mathbf{x}_0 + \sum_{i=0}^{k} \mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i \right) \\
&\sim N\left(\mathbf{L}(k,0)\mathbf{x}_0^b + \sum_{i=0}^{k} \mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i, \mathbf{L}(k,0)\mathbf{B}_0\mathbf{L}^T(k,0) \right) \qquad (83)
\end{aligned}$$

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

$$p(\mathbf{w}_1, \cdots, \mathbf{w}_l) = p_{\mathbf{w}_1}(\mathbf{w}_1) \cdots p_{\mathbf{w}_l}(\mathbf{w}_l) \qquad (84)$$

Substituting all these density functions into (76), we obtain

$$\begin{aligned}
& p(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l | \mathbf{y}_1, \cdots, \mathbf{y}_l) \\
& \quad = c' \exp\left\{ -\frac{1}{2} \left(\mathbf{x}_k - [\mathbf{L}(k,0)\mathbf{x}_0^b + \sum_{i=0}^{k} \mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i] \right)^T \right.
\end{aligned}$$

$$
\begin{aligned}
& \mathbf{L}^{-T}(k,0)\mathbf{B}_0^{-1}\mathbf{L}^{-1}(k,0)\left(\mathbf{x}_k - [\mathbf{L}(k,0)\mathbf{x}_0^b + \sum_{i=0}^{k}\mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i]\right)^T \\
-& \frac{1}{2}\sum_{j=1}^{k}\left(\mathbf{y}_j - \mathbf{h}_j\mathbf{L}^{-1}(k,j)\left[\mathbf{x}_k - \sum_{i=j}^{k}\mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i\right]\right)^T \\
& \mathbf{R}_j^{-1}\left(\mathbf{y}_j - \mathbf{h}_j\mathbf{L}^{-1}(k,j)\left[\mathbf{x}_k - \sum_{i=j}^{k}\mathbf{L}(k,i)\Gamma(i-1)\mathbf{w}_i\right]\right) \\
-& \frac{1}{2}\sum_{j=k+1}^{l}\left(\mathbf{y}_j - \mathbf{h}_j\left[\mathbf{L}(j,k)\mathbf{x}_k + \sum_{i=k}^{j}\mathbf{L}(i,k)\Gamma(i-1)\mathbf{w}_i\right]\right)^T \\
& \mathbf{R}_j^{-1}\left(\mathbf{y}_j - \mathbf{h}_j\left[\mathbf{L}(j,k)\mathbf{x}_k + \sum_{i=k}^{j}\mathbf{L}(i,k)\Gamma(i-1)\mathbf{w}_i\right]\right) \\
-& \sum_{j=1}^{l}\mathbf{w}_j^T\mathbf{Q}_j^{-1}\mathbf{w}_j\Bigg\}, \qquad (85)
\end{aligned}
$$

where c' is independent of $\{\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l\}$. Thus, we can obtain that $\{\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l\}$ are jointly Gaussian as $\{\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l\}$.

We obtain the conditional mean

$$
\hat{\mathbf{x}}_k = \int\cdots\int \mathbf{x}_k p(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l|\mathbf{y}_1, \cdots, \mathbf{y}_l)d\mathbf{x}_k d\mathbf{w}_1\cdots d\mathbf{w}_l. \qquad (86)
$$

This conditional mean $\hat{\mathbf{x}}_k$ is also the maximum likelihood estimate due to the fact that $p(\mathbf{x}_k|\mathbf{y}_1, \cdots, \mathbf{y}_l)$ is Gaussian.

Substituting (38) and (85) into (86), we obtain

$$
\begin{aligned}
\hat{\mathbf{x}}_k &= \int\cdots\int \mathbf{L}(k,0)\mathbf{x}_0 p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l|\mathbf{y}_1, \cdots, \mathbf{y}_l)d\mathbf{x}_0 d\mathbf{w}_1\cdots d\mathbf{w}_l \\
&+ \int\cdots\int \sum_{j=1}^{k}\mathbf{L}(k,j)\Gamma(j-1)\mathbf{w}_j d\mathbf{x}_0 d\mathbf{w}_1\cdots d\mathbf{w}_l. \qquad (87)
\end{aligned}
$$

Here we have used the fact that $p(\mathbf{x}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l|\mathbf{Y}_l)$ becomes identical to $p(\mathbf{x}_0, \mathbf{w}_1, \cdots, \mathbf{w}_l|\mathbf{Y}_l)$ in (75) when (38) is substituted into (85). In the derivation of (87), we have used the change of variable in the integrals and the relation of the probability density for a function of random variables.

From (87) it follows that

$$\hat{\mathbf{x}}_k = \mathbf{L}(k,0)\hat{\mathbf{x}}_0 + \sum_{j=1}^{k} \mathbf{L}(k,j)\Gamma(j-1)\hat{\mathbf{w}}_j \tag{88}$$

We thus conclude that the 4D-Var solution is optimal not only with respect to the model trajectory consisting of $(\mathbf{x}_0, \cdots, \mathbf{x}_l)$ but also with respect to $\mathbf{x}_k$ at a single observation time though the model is not perfect. Obviously, this property of the 4D-Var solution is associated with the consistent property of the solution on the time interval. Thus, we may call this optimality property as optimality consistence.

In association with its additive property, 4D-Var possess an additive optimality property. When observations are separated into subsets, 4D-Var is carried out in the way described in section 3. The final 4D-Var solution is optimal as all the observations are processed simultaneously.

5 Equivalence of 4D-Var to the Kalman filter and smoother

Prior to the discussion of the equivalence, we further present an optimal property of 4D-Var, which we call an optimal transferability property. 4D-Var can only be carried out over a limited time window. Here we consider a time window with observations

$$\{\mathbf{y}_1, \cdots, \mathbf{y}_l, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}\}. \tag{89}$$

The observation is separated into two batches, namely,

$$\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}\}. \tag{90}$$

We solve two 4D-Var problems: the first one with observations $\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$; the second one with $\{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}\}$, and with the background field $\hat{\mathbf{x}}_l$ provided by the previous batch. Then the 4D-Var solution $\{\hat{\mathbf{x}}_l, \cdots, \hat{\mathbf{x}}_{l+m}\}$ yielded by the second batch is optimal as in the

4D-Var where all data in the two hatches are processed simultaneously, namely, $\hat{\mathbf{x}}_k$ maximizes $p(\mathbf{x}_k|Y_{l+m})$ for $(l \leq k \leq l+m)$. This property shows that information from one batch can be fully transferred to the next via the background term.

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

$$
\begin{aligned}
& p(\mathbf{x}_l, \cdots, \mathbf{x}_{l+m}|\mathbf{Y}_{l+m}) \\
&= \frac{p(\mathbf{x}_l, \cdots, \mathbf{x}_{l+m}, \mathbf{Y}_l, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m})}{p(\mathbf{Y}_{l+m})} \\
&= \frac{p(\mathbf{Y}_l)}{p(\mathbf{Y}_{l+m})} p(\mathbf{x}_l, \cdots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}|\mathbf{Y}_l) \\
&= \frac{p(\mathbf{Y}_l)}{p(\mathbf{Y}_{l+m})} p(\mathbf{x}_l|\mathbf{Y}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}) p(\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}|\mathbf{x}_l, \mathbf{Y}_l) \\
&= \frac{p(\mathbf{Y}_l)}{p(\mathbf{Y}_{l+m})} p(\mathbf{x}_l|\mathbf{Y}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}) \\
&\quad p(\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}|\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{x}_l, \mathbf{Y}_l) p(\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}|\mathbf{x}_l, \mathbf{Y}_l) \qquad (91)
\end{aligned}
$$

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

$$
p(\mathbf{x}_l|\mathbf{Y}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}) = p(\mathbf{x}_l|\mathbf{Y}_l) \qquad (92)
$$

In view of (4), we get

$$
\begin{aligned}
& p(\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}|\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}, \mathbf{x}_l, \mathbf{Y}_l) \\
&= p(\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}|\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}) \\
&= p_{\mathbf{v}_{l+1}}(\mathbf{y}_{l+1} - \mathbf{h}_{l+1}\mathbf{x}_{l+1}) \cdots p_{\mathbf{v}_{l+m}}(\mathbf{y}_{l+m} - \mathbf{h}_{l+m}\mathbf{x}_{l+m}) \qquad (93)
\end{aligned}
$$

Note that $\mathbf{w}_k$ is independent of Y_l for $k > l$. Again, since the sequence $\mathbf{x}_l$ is Markov, we obtain

$$
p(\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m}|\mathbf{x}_l, \mathbf{Y}_l)
$$

$$
\begin{aligned}
&= p(\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+m} | \mathbf{x}_l) \\
&= p(\mathbf{x}_{l+m} | \mathbf{x}_{l+m-1}, \cdots p(\mathbf{x}_{l+1}), \mathbf{x}_l) \\
&= p_{\mathbf{w}_{l+1}}(\mathbf{x}_{l+1} - \mathbf{L}(l+1, l)\mathbf{x}_l) \cdots p_{\mathbf{w}_{l+m}}(\mathbf{x}_{l+m} - \mathbf{L}(l+m, l+m-1)\mathbf{x}_{l+m-1}) \qquad (94)
\end{aligned}
$$

Substituting (92), (93) and (94) into (91), we thus obtain that maximizing $p(\mathbf{x}_l, \cdots, \mathbf{x}_{l+m} | \mathbf{Y}_{l+m})$ is equivalent to minimizing the cost function

$$
\begin{aligned}
J &= \frac{1}{2}[\mathbf{x}_l - \hat{\mathbf{x}}_l]^T (\mathbf{P}_l^a)^{-1} [\mathbf{x}_l - \hat{\mathbf{x}}_l] + \frac{1}{2} \sum_{k=l+1}^{l+m} (\mathbf{h}_k \mathbf{L}(k, l)\mathbf{x}_l - \mathbf{y}_k)^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{L}(k, l)\mathbf{x}_l - \mathbf{y}_k) \\
&+ \frac{1}{2} \sum_{k=l+1}^{l+m} \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k, \qquad (95)
\end{aligned}
$$

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

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) give 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.

We here further show why such equivalence occurs. For one iteration of the Kalman filter, we have at step $k-1$

$$
\mathbf{x}_{k-1} \sim N(\hat{\mathbf{x}}_{k-1}, \mathbf{P}_{k-1}^a). \qquad (96)
$$

We define a cost function corresponding to (95)

$$
J_k = \frac{1}{2}(\mathbf{y}_k - \mathbf{h}_k \mathbf{x}_k)^T \mathbf{R}_k^{-1} (\mathbf{y}_k - \mathbf{h}_k \mathbf{x}_k) + \mathbf{w}_k^T \mathbf{Q}_k^{-1} \mathbf{w}_k + (\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1})^T \mathbf{B}_{k-1}^{-1} (\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1}). \qquad (97)
$$

Since

$$\mathbf{x}_{k-1} = \mathbf{L}^{-1}(k, k-1)(\mathbf{x}_k - \Gamma(k-1)\mathbf{w}_k), \tag{98}$$

the cost function (97) becomes

$$\begin{aligned} J_k &= \frac{1}{2}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k)^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{x}_k) + \mathbf{w}_k^T\mathbf{Q}_k^{-1}\mathbf{w}_k \\ &+ [\mathbf{L}^{-1}(k, k-1)(\mathbf{x}_k - \Gamma(k-1)\mathbf{w}_k) - \hat{\mathbf{x}}_{k-1}]^T \\ &\quad \mathbf{B}_{k-1}^{-1}[\mathbf{L}^{-1}(k, k-1)(\mathbf{x}_k - \Gamma(k-1)\mathbf{w}_k) - \hat{\mathbf{x}}_{k-1}] \end{aligned} \tag{99}$$

Let

$$\frac{\partial J_k}{\partial \mathbf{x}_k} = 0, \tag{100}$$

$$\frac{\partial J_k}{\partial \mathbf{w}_k} = 0. \tag{101}$$

We obtain

$$\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{h}_k\mathbf{x}_k - \mathbf{y}_k) + \mathbf{L}^{-1}(k, k-1)\mathbf{B}_{k-1}^{-1}[\mathbf{L}^{-1}(k, k-1)(\mathbf{x}_k - \Gamma(k-1)\mathbf{w}_k) - \hat{\mathbf{x}}_{k-1}] = 0 \tag{102}$$

$$\begin{aligned} &[\mathbf{Q}_k^{-1} + \Gamma^T(k-1)\mathbf{L}^{-T}(k, k-1)\mathbf{B}_{k-1}^{-1}\mathbf{L}^{-1}(k, k-1)\Gamma(k-1)]\mathbf{w}_k \\ &+\Gamma^T(k-1)\mathbf{L}^{-1}(k, k-1)\mathbf{B}_{k-1}^{-1}[\hat{\mathbf{x}}_{k-1} - \mathbf{L}^{-1}(k, k-1)\mathbf{x}_k = 0 \end{aligned} \tag{103}$$

Eliminating $\mathbf{w}_k$, we obtain

$$\begin{aligned} &[\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k + \mathbf{L}^{-T}(k, k-1)\mathbf{B}_{k-1}^{-1}\mathbf{L}^{-1}(k, k-1) \\ &-\mathbf{L}^{-T}(k, k-1)\mathbf{B}_{k-1}^{-1}\mathbf{L}^{-1}(k, k-1)\Gamma(k-1)\mathbf{H}_{Q_k}^{-1}\Gamma^T(k-1) \\ &\mathbf{L}^{-T}(k, k-1)\mathbf{B}_{k-1}^{-1}\mathbf{L}^{-1}(k, k-1)](\mathbf{x}_k - \mathbf{L}(k, k-1)\hat{\mathbf{x}}_{k-1}) \\ = \ &\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}(k, k-1)\hat{\mathbf{x}}_{k-1}), \end{aligned} \tag{104}$$

where

$$\mathbf{H}_{Q_k} = \mathbf{Q}_k^{-1} + \Gamma^T(k-1)\mathbf{L}^{-T}(k, k-1)\mathbf{B}_{k-1}^{-1}\mathbf{L}^{-1}(k, k-1)\Gamma(k-1). \tag{105}$$

We define

$$\mathbf{P}_k^f = \mathbf{L}(k,k-1)\mathbf{B}_{k-1}\mathbf{L}^T(k,k-1) + \Gamma(k-1)\mathbf{Q}_k\Gamma^T(k-1). \tag{106}$$

Then we have

$$\mathbf{H}_{Q_k}^{-1} = \mathbf{Q}_k - \mathbf{Q}_k\Gamma^T(k-1)(\mathbf{P}_k^f)^{-1}\Gamma(k-1)\mathbf{Q}_k \tag{107}$$

By (107), we further have

$$\begin{aligned}
\Gamma(k-1)\mathbf{H}_Q^{-1} &= [I - \Gamma(k-1)\mathbf{Q}_k\Gamma^T(k-1)(\mathbf{P}_k^f)^{-1}]\Gamma(k-1)\mathbf{Q}_k \\
&= \mathbf{L}(k,k-1)\mathbf{B}_{k-1}\mathbf{L}^T(k,k-1)(\mathbf{P}_k^f)^{-1}\Gamma(k-1)\mathbf{Q}_k.
\end{aligned} \tag{108}$$

The substitution of (108) into (104) yields

$$[\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k + (\mathbf{P}_k^f)^{-1}](\mathbf{x}_k - \mathbf{L}(k,k-1)\hat{\mathbf{x}}_{k-1}) = \mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}(k,k-1)\hat{\mathbf{x}}_{k-1}) \tag{109}$$

Thus the minimization solution is

$$\hat{\mathbf{x}}_k = \mathbf{L}(k,k-1)\hat{\mathbf{x}}_{k-1} + [\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k + (\mathbf{P}_k^f)^{-1}]^{-1}\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}(k,k-1)\hat{\mathbf{x}}_{k-1}). \tag{110}$$

This solution is identical to the Kalman filter solution. A similar derivation can be found in Bryson and Ho (1975). Based on the property of 4D-Var presented above, 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, p166). 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 $0 \leq k \leq l$. 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 applicable in meteorology, thus it will be not discussed in the following.

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 $(\mathbf{x}_0, \cdots, \mathbf{x}_l)$ but also with respect to $\mathbf{x}_k$ at a single observation time. One single evaluation of 4D-Var yields all optimal estimates $(\hat{\mathbf{x}}_0, \cdots, \hat{\mathbf{x}}_l)$.

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 can not be directly applied. This is the case when we are facing 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 reanalysis projects (e.g., Gibson et al. 1997; Kalnay et al. 1996; Schubert and Rood 1995). 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 the fixed-lag smoothing should be the ultimate goal of reanalysis efforts. Thus, 4D-Var is practically not equivalent to the fixed-lag smoother.

c A 4D-Var 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). The particular optimality properties and computational efficiency of 4D-Var provide us with various practical choices for combining 4D-Var into a smoother. In what follows, we propose such an approach.

This approach is based on a batch processing approach. The batch processing approach consists in separating the observations into batches.

$$\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{2l}\}, \cdots, \{\mathbf{y}_{(j-1)l+1}, \cdots, \mathbf{y}_{jl}\}, \cdots, \tag{111}$$

and solving the standard 4D-Var problem separately for each batch of observations (Jazwinski 1971).

It has been known that the Kalman filter is optimal not only at every time step, but over the entire interval over which data are provided (e.g., Kalman 1960; Ghil and Malanotte-Rizzoli 1991). Due to the optimality transferability property of 4D-Var, the batch processing possesses the optimality similar to the Kalman filter. In the maximum likelihood terminology, the 4D-Var solution $\hat{\mathbf{x}}_{(j-1)l+k}$ maximizes

$$p(\mathbf{x}_{(j-1)l+k}|\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_j) \tag{112}$$

where $1 \leq k \leq l$, and $\mathbf{Y}_j = \{\mathbf{y}_{(j-1)l+1}, \cdots, \mathbf{y}_{jl}\}$. This is the necessary result of a linear and Markov process.

Such an optimalilty transferability 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). We are seeking the optimal $\hat{\mathbf{x}}_k$ that maximizes $p(\mathbf{x}_k|\mathbf{y}_1, \cdots, y_k, \cdots, y_{k+N})$ for all k with N fixed. The constructive procedure is outlined as follows:

1. We separate the observations into a sequence of batches ($S0$)

$$\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{2l}\}, \cdots, \tag{113}$$

   and to solve the standard 4D-Var problem separately for each batch of observations.

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 (Sm),

$$\{\mathbf{y}_{1+md}, \cdots, \mathbf{y}_{l+md}\}, \{\mathbf{y}_{l+md+1}, \cdots, \mathbf{y}_{2l+md}\}, \cdots, \tag{114}$$

where $m = 1, \cdots, l/d \leq l$. 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 analyes.

3. Using these $M+1$ assimilated analysis sequences, we can construct a time-consistent climatological data sequence by picking one single assimilation data 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 $p(\mathbf{x}_k|\mathbf{y}_1, \cdots, y_k, \cdots, y_{k+N})$ for all k.

The assimilated data sequence produced by one sequence of batches defined in (111) is not consistent in time. This method 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, that is, all observations must be used only once for the estimation of one variable. To comply with this principle, batch processing seems to be the only choice.

This 4D-Var fixed-lag smoother requires M-batch processings. However, the computational overhead must not necessarily increase M fold. Though M-batch processings are carried out independently, the information generated by one batch processing can be used by another batch processing. This observation leads to two techniques that may be used jointly to accelerate the minimization procedure, and thus reduce the computational overhead.

First, the 4D-Var in the sequence $m-1$ provides the initial guess to the 4D-Var in the sequence m. This technique is the same as the one used in quasi-continuous variational data assimilation due to Jarvinen 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 sequence $m-1$ provides the Hessian information for the 4D-Var in the sequence m. 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 and Navon 1999).

6 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 was currently performed. Also, 4D-Var allows to process 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 is 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 sense in the linear context even for models with errors. The equivalence has been discussed for a "sweep solution" by Bryson and Ho (1975), Bennett and Budgell (1989) and Menard and Daley (1996).

Difficulties arise when 4D-Var is used to produce a consistent and long sequence of assimilation data. This is the major purpose of data reanalysis (Gibson et al. 1997; Kalnay et al. 1996; Schubert and Rood 1995). 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 evidently applicable to this task.

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*. Along with computational efficiency of 4D-Var, we thus have various practical choices for combining 4D-Var into a smoother that produces a consistent and long sequence of climatological data. Here 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 a promise for practical use.

However, in order to implement the equivalence of 4D-Var to the Kalman filter/smoother, two requirements must be satisfied: all statistics of covariances used in 4D-Var should be identical to those used in the Kalman filter and smoother; 4D-Var produces the exact minimum solution of the problem. Unfortunately, this is not the case in the reality. The sequential method produces the optimal solution along with the evolution of covariances. 4D-Var does not generate the time evolution of covariances. 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 numerical weather prediction, but it constitutes a serious concern when 4D-Var is used for producing climatological data as in the proposed multi-processing 4D-Var smoother.

Thus, in order to render 4D-Var truly equivalent to the sequential method, the procedure of evaluating covariances in 4D-Var must be developed independently. In this domain, we have seen significant progress. Todling et al. (1998) have summaried and tested major simplified techniques for the computation of the time evolution of error covariances in simple models. A simplified Kalman filter based on a partial eigendecomposition method (Todling et al. 1998) has been used to improve the estimation of covariances in 4D-Var (Fisher 1998; Rabier et al. 1997). This method presents encouraging results and is expected to move toward operational implementation. A difficulty in this method is that it requires an explicit expression of the analysis error covariance. As such, it is not yet clear if this method may also be applicable when model errors are considered, since in this case it is difficult to obtain an explicit expression of the analysis error covariance. An alternative is the ensemble method due to Evensen (1994). The statistics of covariances is estimated based on the ensemble (e.g., Evensen 1994, 1997; Houtekamer and Mitchell 1998). The ensemble method is computationally efficient and applicable to both imperfect and nonlinear models. The ensemble method may be more reliably used if the estimation of the error in the covariance generated from the ensemble is provided.

Both the sequential method and variational method face difficulties in dealing with model errors. A serious difficulty consists in how to define the statistics of model errors, though model bias seems manageable (e.g. Derber 1989; Dee 1998). In 4D-Var, dealing with model errors requires an enormous increase in computational overhead as shown in section 3. A tremendous simplification is necessary to render 4D-Var viable in an operational environment, as done in the representer method (e.g., Bennett 1992) or the general weak constraint method (D. Zupanski 1997).

We have detailed the optimality of 4D-Var solutions and the equivalence between the variational method and sequential method, and discussed possible techniques for implementing such equivalence in real-life applications. We would like to emphasize that we have made no attempt to determine which of the two methods performs better. In practice, the performance of each method strongly depends on model performance, quantity of observations, use of assimilated data, and the suboptimality implied in the implementation along with specific computational burdens. This is rather a practical problem, much more than a theoretical problem.

Acknowledgments

We acknowledge the support from the NSF grant number ATM-9731472 managed by Dr. Pamela Stephens whom we would like to thank for her support and from the Supercomputer Computations Research Institute, which is partially funded by the Department of Energy through Contract DE-FG05-85ER250000.

A Appendix

We prove here that the 4D-Var solution for observations

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

is identical to the batch-processing solution for

$$\{\mathbf{y}_1, \cdots, \mathbf{y}_l\}, \{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}\}, \tag{A.2}$$

at the observation time corresponding to $\{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_{l+m}\}$.

A perfect model is considered. The conclusion holds for a model with errors under Gaussian distributions. We may write the cost function associated with the second batch in (A.2)

$$J(\mathbf{x}_l) = \frac{1}{2}[\mathbf{x}_l - \hat{\mathbf{x}}_l]^T (\mathbf{P}_l^a)^{-1} [\mathbf{x}_l - \hat{\mathbf{x}}_l] + \frac{1}{2} \sum_{k=l+1}^{l+m} (\mathbf{h}_k \mathbf{L}(k,l)\mathbf{x}_l - \mathbf{y}_k)^T \mathbf{R}_k^{-1} (\mathbf{h}_k \mathbf{L}(k,l)\mathbf{x}_l - \mathbf{y}_k). \tag{A.3}$$

where $\hat{\mathbf{x}}_l$ is the 4D-Var solution associated with the first batch in (A.2), given by (20), i.e.,

$$\hat{\mathbf{x}}_l = \mathbf{L}(l,0)[\mathbf{x}_0^b + (\mathbf{H}_{0,l})^{-1}\sum_{k=1}^{l}\mathbf{L}(k,0)^T\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{y}_k^b)], \tag{A.4}$$

where

$$\mathbf{y}_k^b = \mathbf{h}_k\mathbf{L}(k,0)\mathbf{x}_0^b. \tag{A.5}$$

$(\mathbf{P}_l^a)^{-1}$ is the inverse of the analysis error covariance matrix associated with $\hat{\mathbf{x}}_l$, given by

$$(\mathbf{P}_l^a)^{-1} = \mathbf{L}^{-T}(l,0)\mathbf{H}_{0,l}\mathbf{L}^{-1}(l,0) \tag{A.6}$$

Following the procedure in section 3, we obtain the 4D-Var solution

$$\hat{\hat{\mathbf{x}}}_l = \hat{\mathbf{x}}_l + (\mathbf{H}_{l,l+m})^{-1}\sum_{k=l+1}^{l+m}\mathbf{L}^T(k,l)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}(k,l)\hat{\mathbf{x}}_l), \tag{A.7}$$

where

$$\mathbf{H}_{l,l+m} = (\mathbf{P}_l^a)^{-1} + \sum_{k=l+1}^{l+m}\mathbf{L}^T(k,l)\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k,l). \tag{A.8}$$

The substitution of (A.6) into (A.8) yields

$$\mathbf{H}_{l,l+m} = \mathbf{L}^{-T}(l,0)\mathbf{H}_{0,l+m}\mathbf{L}^{-1}(l,0) \tag{A.9}$$

where

$$\mathbf{H}_{0,l+m} = \mathbf{B}_0^{-1} + \sum_{k=1}^{l+m}\mathbf{L}(k,0)^T\mathbf{h}_k^T R_k^{-1}\mathbf{h}_k\mathbf{L}(k,0) \tag{A.10}$$

is the Hessian associated with 4D-Var for all the observation as a whole.

Substituting (A.4) and (A.9) into (A.7), we obtain

$$\begin{aligned}
\hat{\hat{\mathbf{x}}}_l &= \mathbf{L}(l,0)\left\{\mathbf{x}_0^b + \mathbf{H}_{0,l}^{-1}\sum_{k=1}^{l}\mathbf{L}(k,0)^T\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{y}_k^b)\right. \qquad \text{(A.11)}\\
&\quad \left. + \mathbf{H}_{0,l+m}^{-1}\mathbf{L}^T(0,l)\sum_{k=l+1}^{l+m}\mathbf{L}^T(k,l)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{h}_k\mathbf{L}(k,l)\hat{\mathbf{x}}_l)\right\}\\
&= \mathbf{L}(l,0)\left\{\mathbf{x}_0^b + \mathbf{H}_{0,l}^{-1}\sum_{k=1}^{l}\mathbf{L}(k,0)^T\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{y}_k^b)\right.
\end{aligned}$$

$$
+\mathbf{H}_{0,l+m}^{-1}\left[\sum_{k=1}^{l+m}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{y}_k^b)-\sum_{k=1}^{l}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{y}_k^b)\right.
$$

$$
\left.\left.-\sum_{k=l+1}^{l+m}\mathbf{L}^T(k,l)\mathbf{h}_k^T\mathbf{R}_k^{-1}\mathbf{h}_k\mathbf{L}(k,l)(\mathbf{H}_{0,l})^{-1}\sum_{k=1}^{l}\mathbf{L}(k,0)^T\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{y}_k^b)\right]\right\}
$$

Since

$$
\mathbf{H}_{0,l+m}=\mathbf{H}_{0,l}+\sum_{k=l+1}^{l+m}\mathbf{L}(k,0)^T\mathbf{h}_k^T R_k^{-1}\mathbf{h}_k\mathbf{L}(k,0), \tag{A.12}
$$

we finally obtain

$$
\hat{\hat{\mathbf{x}}}_l=\mathbf{L}(l,0)\left\{\mathbf{x}_0^b+\mathbf{H}_{0,l+m}^{-1}\sum_{k=1}^{l+m}\mathbf{L}^T(k,0)\mathbf{h}_k^T\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{y}_k^b)\right\} \tag{A.13}
$$

From (13) and (26), we conclude that $\hat{\hat{\mathbf{x}}}_l$ is identical to the 4D-Var solution at the observation time t_l when all $l+m$ observations are processed simultaneously. Due to the optimality of 4D-Var, this conclusion holds for every observation time level t_k ($k=l,\cdots,l+m$).

REFERENCES

Anderson, B.D.O., and Moore, J.B., 1979: *Optimal filtering*, Prentice-Hall, 357pp.

Bennett, A.F., 1997: *Inverse methods in physical oceanography.* Cambridge Univ. Press, Cambridge, UK, 346pp.

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

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

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. Wea. Rev.*, **120**, 1661-1674.

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

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

Cohn, S.E., Todling, R., and Zhu, Y., 1998: *Office Note*, NASA/GSFC Data Assimilation Office, 66pp.

Courtier, P., 1997: Variational methods. *J. Meteor. Soc Japan*, **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., Thepaut, 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

Daley R. 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.

Evensen, G., 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Weather. Rev.*, **125**, 1342-1354.

Fisher, M., 1998: Development of a simplified Kalman filter. *Technical memorandum No. 260*, ECMWF.

Ghil, M., 1989: Meteorological data assimilation for oceanographer. Part I. Description and theoretical framework. *Dym. Atmos. Oceans*, **13**, 171-218.

Ghil, M., Cohn, S.E., Tavantzis, J., Bube, K., and Isaacson, E., 1981: Application of estimation theory to numerical weather prediction. *Dynamical Meteorological: Data Assimilation Methods*, L. Bengtsson, M. Gill and E. Kallen, Eds., Spring-Verlag, 139-224.

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

Ghil, M., 1997: Advances in sequential estimation for atmospheric and oceanic flows. *J. Meteor. Soc Japan*, **75, 1B**, 289-304.

Gibson, J.K., Hernandez, A., Kallberg, P., Nomura, A., Serrano, E., and Uppala, S., 1996: Current status of the ECMWF re-analysis (ERA) project. *ECMWF Workshop proceedings,* *Fifth Workshop on Meteorological Operational Systems*, 283-290.

Houtekamer, P.L., and Mitchell, H.L., 1998: Data assimilation using an Ensemble Kalman filter technique. *Mon. Weather. Rev.*, **126**, 796-811.

Jarvinen, H. Thepaut, 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

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

Klinker, E., Rabier, F., and Mahfouf, J.-F., 1998: The ECMWF operational implementation of four-dimensional variational assimilation. Part III: Experimental results and diagnostics with operational configuration. *Q. J. R. Meteorol. Soc.*, submitted.

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., 1999: Performance of 4D-Var with different strategies on the use of adjoint physics with the FSU Global Spectral Model. *Mon. Weather. Rev.*, in press.

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

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

Mahfouf, J.-F., and Rabier, F., 1998: The ECMWF operational implementation of four-dimensional variational assimilation. Part II: Experimental results with improved physics. *Q. J. R. Meteorol. Soc.*, submitted.

Navon, I. M., Zou, X., Derber, J., and Sela, J., 1992b: 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., 1998a: The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, submitted.

Rabier, F., Mahfouf, J.-F., Fisher, M., Jarvinen, H., Simmons, A., Andersson, E., Bouttier, F., Courtier, P., Hamrud, H., Haseler, J. Hollingsworth, A., Isaksen, L., Klinker, P., Saarinen, S., Temperton, C., Thepaut, J.-N., Unden, P., and Vasiljevic, D., 1997: Recent experimentation on 4D-Var and first results from a simplified Kalman filter. *ECMWF Research Department Tech. Memo. No. 240*, ECMWF, Reading, UK, pp42.

Rabier, F., Thepaut, J.-N., and Courtier, P., 1998b: 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: Proceeding of the workshop on the GEOS-1 five-year assimilation. *NASA Tech. Memo. 104606*, **Vol. 7**, 201pp.

Talagrand, O., and Courtier, P., 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I. Theory. *Q. J. R. Meteorol. Soc.*, **113**, 1131-1328.

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.

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

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 system. *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.