Variational Analysis Using Spatial Filters

XIANG-YU HUANG

Danish Meteorological Institute, Copenhagen, Denmark

(Manuscript received 7 January 1999, in final form 15 October 1999)

ABSTRACT

In this paper the standard variational analysis scheme is modified, through a simple transform, to avoid the inversion of the background error covariance matrix. A close inspection of the modified scheme reveals that it is possible to use a filter to replace the multiplication of the covariance matrix. A variational analysis scheme using a filter is then formulated, which does not explicitly involve the covariance matrix. The modified scheme and the filter scheme have the advantage of avoiding the inversion or any usage of the large matrix for analyses using gridpoint representation. To illustrate the use of these schemes, a small-sized and a more realistic analysis problem is considered using real temperature observations. It is found that both the modified scheme and the filter scheme work well. Compared to the standard and modified schemes the storage and computational requirements of the filter scheme can be reduced by several orders of magnitude for realistic atmospheric applications.

1. Introduction

Objective atmospheric data analysis is widely used in observational studies, numerical weather prediction (NWP), and for verifying new models and theories. In many analysis schemes, observations are optimally used together with a background field in a statistical sense. At most NWP centers, a particular implementation of an Optimal Interpolation (OI) (Eliassen 1954; Gandin 1963) or a variational data assimilation (VAR) (Lorenc 1986) is used. In the OI and VAR formulations, the observation and background error covariance matrices, **R** and **B**, are required. The sizes of **R** and **B** are of order $10^5 \times 10^5$ to $10^7 \times 10^7$ for realistic atmospheric applications (Courtier 1997). Special considerations are needed in handling of **R** and **B** in each implementation, especially when their inverse is needed.

In most OI implementations **B** is computed at the observation locations, $\mathbf{HBH}^T$, where **H** is a linear observation operator that maps variables from model grids to observation locations. The inverse of $\mathbf{HBH}^T + \mathbf{R}$ is required by OI. Data selection algorithms, for example, the box structures in the OI implementation at the European Centre for Medium-Range Weather Forecasts (the ECMWF OI was replaced by a variational scheme in 1997), are needed to reduce $\mathbf{HBH}^T + \mathbf{R}$ to a manageable size.

In most VAR implementations there is no OI-type data selection. The most common assumption about **R** is that the observation errors are not correlated (for most variables), which makes **R** diagonal (e.g., Courtier et al. 1998). Using spectral representations, **B** is also diagonal (Parrish and Derber 1992; Courtier et al. 1998; Gustafsson et al. 1999). However, if an implementation of the VAR scheme using a gridpoint formulation is desirable, the inversion of **B** is nontrivial. There are at least two potential problems, both due to the size of **B**. The first is how to invert **B**. This is impossible for a realistic atmospheric implementation and avoided by different transforms (Lorenc 1988; Derber and Rosatti 1989). The second is how to handle the storage and computation related to **B**. This is a difficult issue in operational implementations.

In this paper the first problem is addressed by modifying the standard variational analysis scheme, so that the inversion is not needed. The second problem is then addressed by formulating a new scheme that uses a filter to replace the use of the matrix so that **B** itself does not enter the analysis. The modified scheme has been proposed by Lorenc (1988) and Derber and Rosati (1989) to speed up the convergence of descent algorithms such as the steepest descent and conjugate-gradient methods. Here this scheme is reconsidered with the emphasis on the possibility of avoiding the matrix inversion and producing further simplifications. The latter leads to the new filter scheme. The basic idea of the new filter scheme, using a filter to model the full background error covariance matrix, is common in a number of existing analysis schemes using spatial filters (Lorenc 1992; Dé-

Corresponding author address: Dr. Xiang-Yu Huang, Danish Meteorological Institute, Lyngbyvej 100, DK-2100 Copenhagen Ø, Denmark.
E-mail: xyh@dmi.dk

vényi and Benjamin 1998). However, the way the filter is applied in the new scheme makes the scheme more flexible in practical implementations.

This paper is organized as follows. The standard, the modified, and the new schemes are described in section 2 together with comparisons to other related schemes. An illustrative example is given in section 3 using real 2-m temperature observations, followed by another more realistic example in section 4. The conclusions are summarized in section 5.

2. Analysis methods

a. VAR: Variational analysis

Briefly, a three-dimensional variational analysis (3DVAR) can be formulated as the minimization of the following cost function (Lorenc 1986; Courtier 1997):

$$\begin{aligned} J(\mathbf{x}) &= J_b(\mathbf{x}) + J_o(\mathbf{x}) \\ &= \frac{1}{2}\{(\mathbf{x} - \mathbf{x}^b)^{\mathrm{T}}\mathsf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) \\ &\quad + [H(\mathbf{x}) - \mathbf{y}]^{\mathrm{T}}\mathsf{R}^{-1}[H(\mathbf{x}) - \mathbf{y}]\}, \end{aligned}$$

where $\mathbf{x}$ is the analysis vector (on model grid points), $\mathbf{x}^b$ is the background vector (on model grid points), $\mathbf{y}$ is the observation vector (on observation locations), H is the (nonlinear) observation operator that converts model variables to observations, B is the background error covariance matrix, R is the observation error covariance matrix that includes the effects of both instrument and representativeness, $(\cdot)^{\mathrm{T}}$ indicates transpose, and $(\cdot)^{-1}$ indicates inversion. In most practical implementations a penalty term for noise control is included in addition to J_b and J_o, but it is omitted here. The notations closely follow Ide et al. (1997).

Starting with an initial guess field $\mathbf{x}^0$, normally $\mathbf{x}^0 = \mathbf{x}^b$, the final analysis $\mathbf{x}^{\mathrm{VAR}} = \mathbf{x}^{\infty}$ can be obtained using a minimization algorithm. The simplest method is the so-called steepest descent, which can be expressed as

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha\boldsymbol{\nabla}_{\mathbf{x}}J, \tag{1}$$

where the superscript denotes the iteration cycle number, α is a constant ($\alpha > 0$), and $\boldsymbol{\nabla}_{\mathbf{x}}J$ is the gradient of the cost function J with respect to $\mathbf{x}$:

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

where H is the tangent linear operator of H.

Given N model grid points and M observation locations, then $\mathbf{x}$ and $\mathbf{x}^b$ are vectors of length N, $\mathbf{y}$ is a vector of length M, B is a $N \times N$ matrix, and R is a $M \times M$ matrix.

b. VAN: Variational analysis with no inversion of B

To avoid the inversion of B, a new vector $\mathbf{v}$ is introduced, defined as

$$\mathbf{v} = \mathsf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) \tag{3}$$

and the cost function J can now be written as

$$\begin{aligned} J(\mathbf{v}) &= J_b(\mathbf{v}) + J_o(\mathbf{v}) \\ &= \frac{1}{2}\{\mathbf{v}^{\mathrm{T}}\mathsf{B}^{\mathrm{T}}\mathbf{v} + [H(\mathsf{B}\mathbf{v} + \mathbf{x}^b) - \mathbf{y}]^{\mathrm{T}} \\ &\quad \times \mathsf{R}^{-1}[H(\mathsf{B}\mathbf{v} + \mathbf{x}^b) - \mathbf{y}]\}. \end{aligned}$$

The transform from $\mathbf{x}$ to $\mathbf{v}$ using (3) and the redefinition of the cost function with respect to $\mathbf{v}$ were proposed by Lorenc (1988) and Derber and Rosati (1989) to speed up the convergence of descent algorithms such as the steepest descent and conjugate-gradient methods. In this note, VAN is used to provide guidance for the formulation of a new scheme. VAN is also used as a reference when the new scheme is tested.

Using $\mathbf{v}$ as the control variable, the minimization scheme (1) is now rewritten as

$$\mathbf{v}^{n+1} = \mathbf{v}^n - \alpha\boldsymbol{\nabla}_{\mathbf{v}}J, \tag{4}$$

where $\boldsymbol{\nabla}_{\mathbf{v}}$ is the gradient of J with respect to $\mathbf{v}$:

$$\boldsymbol{\nabla}_{\mathbf{v}}J = \mathsf{B}^{\mathrm{T}}\{\mathbf{v} + \mathsf{H}^{\mathrm{T}}\mathsf{R}^{-1}[H(\mathsf{B}\mathbf{v} + \mathbf{x}^b) - \mathbf{y}]\}.$$

Starting with an initial guess field $\mathbf{v}^0$,

$$\mathbf{v}^0 = \mathsf{B}^{-1}(\mathbf{x}^0 - \mathbf{x}^b),$$

the minimum of J is reached at $\mathbf{v}^{\infty}$ and the final analysis $\mathbf{x}^{\mathrm{VAN}}$ becomes

$$\mathbf{x}^{\mathrm{VAN}} = \mathbf{x}^b + \mathsf{B}\mathbf{v}^{\infty}.$$

In this scheme B^{-1} is only needed to provide the initial guess value for the control variable. Consequently, if we assume

$$\mathbf{v}^0 = 0,$$

we have obtained a variational analysis scheme that needs no inversion of B. This assumption is certainly a reasonable one. It simply uses the background vector as the initial analysis vector as done in most existing analysis schemes.

The modified scheme is called VAN, variational analysis with no inversion of B, and is summarized as

$$\begin{aligned} &\mathbf{v}^0 = 0, \\ &\boldsymbol{\nabla}_{\mathbf{v}}J = \mathsf{B}^{\mathrm{T}}\{\mathbf{v} + \mathsf{H}^{\mathrm{T}}\mathsf{R}^{-1}[H(\mathsf{B}\mathbf{v} + \mathbf{x}^b) - \mathbf{y}]\}, \\ &\mathbf{v}^{n+1} = \mathbf{v}^n - \alpha\boldsymbol{\nabla}_{\mathbf{v}}J, \quad \text{and} \\ &\mathbf{x}^{\mathrm{VAN}} = \mathbf{x}^b + \mathsf{B}\mathbf{v}^{\infty}. \end{aligned} \tag{5}$$

The preconditioning has not been discussed. As the optimal preconditioning is the inverse of the Hessian matrix, which is $(\mathsf{B}^{\mathrm{T}} + \mathsf{B}^{\mathrm{T}}\mathsf{H}^{\mathrm{T}}\mathsf{R}^{-1}\mathsf{H}\mathsf{B})^{-1}$ for VAN, some simplifications are needed to really avoid the explicit inversion of B. This issue will be discussed further in the next subsection.

c. VAF: Variational analysis using a filter

In NWP applications, the matrix B ($N \times N \sim 10^{14}$) is too large to be stored in memory by present-day computers and therefore has to be calculated every time it is needed. In VAN, $\mathbf{B}$ is required during each iteration loop. This makes an operational implementation of VAN difficult. However, a close inspection of VAN, in particular of the gradient of the cost function $\nabla_{\mathrm{v}}J$, reveals that $\mathbf{B}$ may be modeled by a spatial filter. The relation between the multiplication of $\mathbf{B}$ and an application of a filter has been discussed by Lorenc (1992). Linear transforms are necessary to construct a filter from $\mathbf{B}$ when a nonuniform grid is used. In this paper, a uniform grid is chosen for simplicity. The generalization of the following scheme to nonuniform grids should be considered in future implementations.

Multiplying a vector $\mathbf{x}$ by the $\mathbf{B}$ matrix, the result vector $\mathbf{x}^*$ has its elements

$$x_i^* = \sum_{j=1}^{N} b_{ij}x_j, \tag{6}$$

where b_{ij} are the elements of $\mathbf{B}$ ($i = 1, N$; $j = 1, N$). In the case of a uniform grid, (6) could be considered as filtering $\mathbf{x}$ with a spatial filter with the filter coefficients b_{ij} and the filter span covering the whole model domain. Therefore the VAN scheme can be written by using a filter to model $\mathbf{B}$. This new scheme is called VAF, variational analysis using a filter, and can be summarized as

$$\begin{aligned}
\mathbf{v}^0 &= 0,\\
\nabla_{\mathrm{v}}J &= \mathcal{G}\{\mathbf{v} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}[H(\mathcal{G}\mathbf{v} + \mathbf{x}^b) - \mathbf{y}]\},\\
\mathbf{v}^{n+1} &= \mathbf{v}^n - \alpha\nabla_{\mathrm{v}}J, \quad \text{and}\\
\mathbf{x}^{\mathrm{VAF}} &= \mathbf{x}^b + \mathcal{G}\mathbf{v}^{\infty},
\end{aligned} \tag{7}$$

where $\mathcal{G}$ is a spatial filter that should be designed based on the a priori knowledge of the covariance matrix $\mathbf{B}$.

For example, consider a two-dimensional univariate problem with homogeneous and isotropic background error correlations. The following Gaussian function could be used to calculate $\mathbf{B}$ (Daley 1991)

$$b_{ij} = \sigma_b^2 \exp\left[-\left(\frac{r_{ij}}{L}\right)^2\right], \tag{8}$$

where σ_b is the standard deviation of the background error, r_{ij} is the distance between model grid points i and j, and L is a length scale that can be determined theoretically or by observations. If the $\mathbf{B}$ matrix is constructed by (8), then the multiplication in (6) is equivalent to applying a Gaussian filter to the vector.

The motivation for using a filter to model $\mathbf{B}$ is to reduce the memory consumption of storing $\mathbf{B}$ and the computational burden caused by multiplying $\mathbf{B}$. Lorenc (1992) and Hayden and Purser (1995) have demonstrated that a multiple iteration recursive filter (no extra memory is required) asymptotically approaches to a Gaussian filter as the iteration tends to infinity. In this paper, a truncated Gaussian filter is used. Although some extra memory is needed to store the filter coefficients, this approach is more flexible. Other forms of $\mathbf{B}$ (Daley 1991) may easily be used to determine the filter coefficients. By using a truncated Gaussian filter, we in fact set many elements in $\mathbf{B}$ to zero. There is no guarantee of positiveness of the modified $\mathbf{B}$. An additional filter windowing is necessary for the convergence of iterations.

Let $N_\lambda \times N_\phi = N$, λ, and ϕ denote coordinates in the west–east and south–north directions, respectively. A two-dimensional filter is chosen for $\mathcal{G}$, with the filter coefficients calculated by (8) and modified by a Lanczos window:

$$\begin{aligned}
\mathbf{x}^* &= \mathcal{G}\mathbf{x},\\
x_i^* &= \sum_{\substack{-0.5I\le(\lambda_j-\lambda_i)/\Delta\lambda\le0.5I\\ -0.5J\le(\phi_j-\phi_i)/\Delta\phi\le0.5J}} g_{ij}x_j,\\
g_{ij} &= w_{ij}\left\{\sigma_b^2 \exp\left[-\left(\frac{r_{ij}}{L_f}\right)^2\right]\right\}, \quad \text{and}\\
w_{ij} &= \frac{\sin[(\lambda_j - \lambda_i)\pi/(0.5I + 1)\Delta\lambda]}{(\lambda_j - \lambda_i)\pi/(0.5I + 1)\Delta\lambda}\\
&\quad \times \frac{\sin[(\phi_j - \phi_i)\pi/(0.5J + 1)\Delta\phi]}{(\phi_j - \phi_i)\pi/(0.5J + 1)\Delta\phi},
\end{aligned} \tag{9}$$

where $\mathbf{x}^*$ is a filtered vector with its elements x_i^*, g_{ij} is the filter coefficient, w_{ij} is the Lanczos window, I and J are the filter orders in the λ and ϕ directions, $\Delta\lambda$ and $\Delta\phi$ are grid distances in the λ and ϕ directions, and L_f is a length scale related to L in b_{ij}. Note that the expression in the curly brackets is the same as b_{ij} except for the length scale. In practice, L is determined by statistics while L_f is based on L and modified due to the filter truncation and window selection.

Due to the symmetry of the filter, the storage of the filter coefficients is $(0.5I + 1)(0.5J + 1)$ and the computation required is $<(I + 1)(J + 1) \times N$ ($<$ is used here because the filter does not take on values outside of the analysis domain). The choice of I and J depends on the characteristics of $\mathbf{B}$. Intuitively, the following approximations should hold (correlations less than 10% can be ignored):

$$\frac{I\Delta\lambda}{2L_f} \approx \frac{J\Delta\phi}{2L_f} \approx \sqrt{2}.$$

Take a typical two-dimensional regional analysis domain, 100 × 100 grid points, with the grid distance $\Delta\lambda = \Delta\phi = 30$ km and the error covariance scale $L = 200$ km. The dimension of $\mathbf{B}$ and the multiplications of $\mathbf{B}$ required by each VAN iteration are $N \times N = 10^8$ and 2×10^8, respectively. The required storage size for the filter is approximately $(0.5I + 1)(0.5J + 1) \approx 10^2$ and the filter multiplications required by each VAF iteration is $2 \times (I + 1)(J + 1) \times N \approx 8 \times 10^6$. (For a VAR scheme using the spectral technique, the storage required for b_{ii}^* and the multiplications in each iteration are $N = 10^4$.)

If the background error at every analysis grid point is correlated with all other grid points, $2(N_\lambda - 1)$ and $2(N_\phi - 1)$ should be chosen for I and J. In this special case, the VAF scheme without the Lanczos window should give exactly the same solution as the VAN scheme, which makes this particular case useful for checking a VAF implementation.

If the background constraint is removed from the VAF scheme, that is, take away the first term $\mathbf{v}$ from the gradient of the cost function $\nabla_\mathbf{v} J$, the scheme becomes a form of successive correction (Bergthorsson and Döös 1955; Cressman 1959; Barnes 1964), which could lead to an analysis fitting observations exactly when the iterations converge. However, if the background contains useful information, as it does for most NWP applications, the constraint should remain. An in-depth discussion on the relations among different analysis schemes can be found in Lorenc (1986).

In a later paper Lorenc (1997) presented a variational assimilation scheme that uses a recursive filter to model $\sqrt{\mathbf{B}}$. He used $\sqrt{\mathbf{B}}$ to define a new control variable $\mathbf{v}$,

$$\sqrt{\mathbf{B}}\mathbf{v} = \mathbf{x} - \mathbf{x}^b,$$

and there is no inverse of $\mathbf{B}$ (or $\sqrt{\mathbf{B}}$) in his proposed scheme. Dévényi and Benjamin (1998) also used $\sqrt{\mathbf{B}}$ to define the new control variable in their 3DVAR scheme. Instead of using recursive filters, an exponential filter was used to model $\sqrt{\mathbf{B}}$. One argument for using $\sqrt{\mathbf{B}}$ when defining the new control variable has been that it provides a better preconditioning for minimization, compared to using $\mathbf{B}$ itself. However, as discussed by Dévényi and Benjamin (1998) this $\sqrt{\mathbf{B}}$ preconditioning is still not good enough for effective operational applications. Some simplified analysis error covariance matrix is needed for the preconditioning. In addition, it is not straightforward to model $\sqrt{\mathbf{B}}$ if $\mathbf{B}$ has complicated structures [e.g. a series of Bessel functions (Lönnberg and Shaw 1987)]. In their latest version of 3DVAR Dévényi and Benjamin also went from modeling $\sqrt{\mathbf{B}}$ to $\mathbf{B}$ (D. Dévényi 1999, personal communication).

d. BLUE: The Best Linear Unbiased Estimation

By setting the gradient of the cost function, (2), to zero and using the linearized observation operator $\mathbf{H}$,

$$\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \mathbf{H}^\mathrm{T}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}) = 0,$$

rearranging the equation,

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

and using the following identity relation,

$$\mathbf{H}^\mathrm{T}\mathbf{R}^{-1}(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R}) = (\mathbf{H}^\mathrm{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})\mathbf{B}\mathbf{H}^\mathrm{T}, \quad \text{or}$$
$$(\mathbf{B}^{-1} + \mathbf{H}^\mathrm{T}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\mathrm{T}\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^\mathrm{T}(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1},$$

the Best Linear Unbiased Estimation (BLUE) can be derived (see, e.g., Courtier 1997):

$$\mathbf{x}^\mathrm{BLUE} = \mathbf{x}^b + \mathbf{B}\mathbf{H}^\mathrm{T}(\mathbf{HBH}^\mathrm{T} + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \quad (10)$$

where $\mathbf{x}^\mathrm{BLUE}$ is the BLUE solution, which in the linear case should be the final solution of variational schemes and therefore can be used as a reference.

The numerical solution for BLUE depends on the accuracy of $(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1}$, which could be problematic, especially when M is large. The size of the matrix $\mathbf{B}\mathbf{H}^\mathrm{T}$ is $N \times M$. The elements of this matrix are the background error covariance between all analysis grid points and observation locations. The size of $\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T}$ is $M \times M$. The elements of this matrix are the background error covariances between all observation locations. Simplifications to $\mathbf{B}\mathbf{H}^\mathrm{T}$ (dimension $N \times M$) and $(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1}$ lead to the widely used OI scheme. In most OI scheme implementations (e.g., Lönnberg and Shaw 1987), different data selection strategies are used to reduce M and N. With the reduced M, both $\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T}$ and $\mathbf{B}\mathbf{H}^\mathrm{T}$ are approximated and directly coded.

e. PSAS: Physical-space Statistical Analysis Scheme

The Physical-space Statistical Analysis Scheme (PSAS) described by Cohn et al. (1998) is another variational analysis scheme avoiding the inversion of $\mathbf{B}$. The BLUE solution can be rewritten as

$$\mathbf{x}^\mathrm{BLUE} = \mathbf{x}^\mathrm{b} + \mathbf{B}\mathbf{H}^\mathrm{T}\mathbf{w} \quad \text{and}$$
$$\mathbf{w} = (\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^b).$$

Instead of solving $\mathbf{w}$ exactly, which needs the inversion of $(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})$ and has the same difficulties in practical applications as discussed in the previous subsection, PSAS obtains $\mathbf{w}$ through minimizing the following cost function:

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

The PSAS solution is given as

$$\mathbf{x}^\mathrm{PSAS} = \mathbf{x}^b + \mathbf{B}\mathbf{H}^\mathrm{T}\mathbf{w}^\infty,$$

where $\mathbf{w}^\infty$ is the minimization result.

Although PSAS does not invert $(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})$, this $M \times M$ matrix could still lead to significant computational expenses. In the implementation of Cohn et al. (1998), simplifications are used to make the matrix sparse, which in a way mimics the OI data selection.

The major difference between PSAS and VAN (VAF) is where the control variable is based. In PSAS, the minimization is performed at the observation locations. The research efforts have been devoted to simplifying $(\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R})$, which is also in the observation space. In VAN (VAF), the minimization is performed at the model grid points. Research efforts are also required to simplify $\mathbf{B}$. As shown in the previous subsections, a simple spatial filter can be used to model $\mathbf{B}$.

3. An illustrative example

To demonstrate the use of the VAN and VAF schemes described in section 2, a small-sized two-dimensional

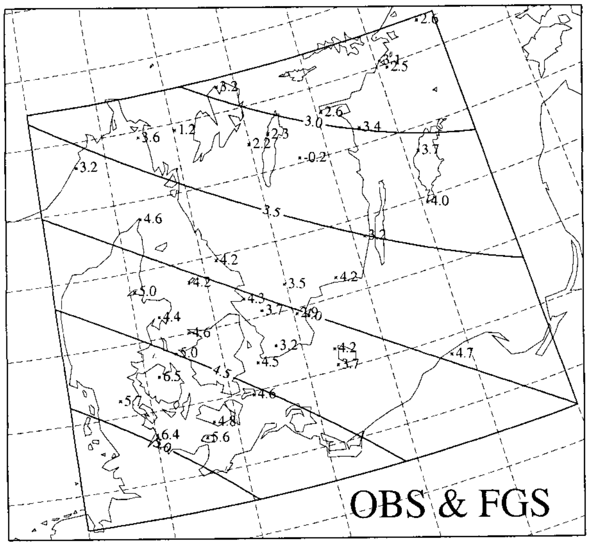

FIG. 1. The 2-m temperature observations (locations are indicated by stars and values are in °C) at 0300 UTC 3 Mar 1992 and the background temperature field (indicated by the contour lines) used in all analyses discussed in section 3.

univariate analysis problem, for which the BLUE solutions are easily obtained as references to evaluate the new schemes, is considered. (A numerical lab, ANALAB, for VAR, VAN, VAF, OI, and PSAS schemes, written in FORTRAN-77, has been developed by the author for educational purposes at the Danish Meteorological Institute. The complete ANALAB is available from the author. ANALAB is based on the OILAB, which has been used in undergraduate courses for many years at the Department of Meteorology, Stockholm University.)

a. Grid, y, x^b, and $\mathbf{H}$

A regular latitude–longitude grid is chosen. The number of analysis grid points is 21 × 21 ($N = 441$). The grid resolution is 0.3° in latitude and 0.6° in longitude. The analysis domain is shown in Fig. 1. The center of the model domain is located at (56.5, 14.0). Using this coordinate, the grid resolution is about 37 km in the east–west direction and 33 km in the north–south direction.

Real 2-m temperature observations are used in this illustrative example. The observations are shown in Fig. 1. They are collected mainly in Denmark and southern Sweden at 0300 UTC 3 March 1992. A total of 38 observations are in the database ($M = 38$). The observation locations are indicated by stars. The observed values in degrees Celsius are printed to the right of their locations.

The background temperature field is also shown in Fig. 1. It is clear that the large-scale features of the temperature field are captured by the background. It is the purpose of the analysis with a smaller error covariance scale to extract the small-scale information in the observations in a statistically optimal way.

The observation operator, $\mathbf{H}$, needed for the following experiments is a simple bilinear interpolation.

b. $\mathbf{R}$, $\mathbf{B}$, *and* $\mathcal{G}$

In the following experiments, the observation errors are assumed noncorrelated. The matrix $\mathbf{R}$ is therefore a diagonal matrix with all diagonal elements equal to the square of the observation standard deviation, σ_o^2, and we choose $\sigma_o = 1$°C.

The background errors are assumed correlated and the Gaussian function, defined in (8), is used to calculate the elements in $\mathbf{B}$. The background standard deviation, σ_b, is also chosen to be 1°C. Using the analysis grid defined earlier, the error covariance between one arbitrary grid point and the neighboring grid points is calculated using (8) and plotted against distance in Fig. 2. In the figure, b200 and b100 are covariances using $L = 200$ km and $L = 100$ km, respectively. In the testing of the VAF code, the 40th-order Gaussian filter without the Lanczos window is used and the filter coefficients, g200_40_nowin, are identical to b200 plotted in Fig. 2.

From the figure, it is clear that the covariance becomes less than 10% only when two grid points are separated by more than 10 grid points when $L = 200$ km and by more than 5 grid points when $L = 100$ km.

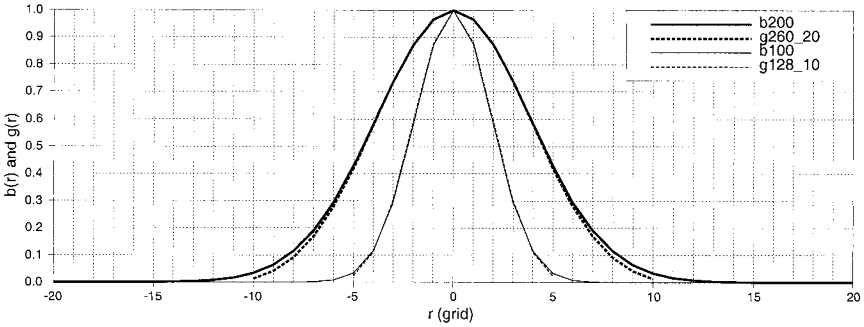

FIG. 2. The background error covariance [$b(r)$] and the Gaussian filter coefficients [$g(r)$], as functions of distance (r).

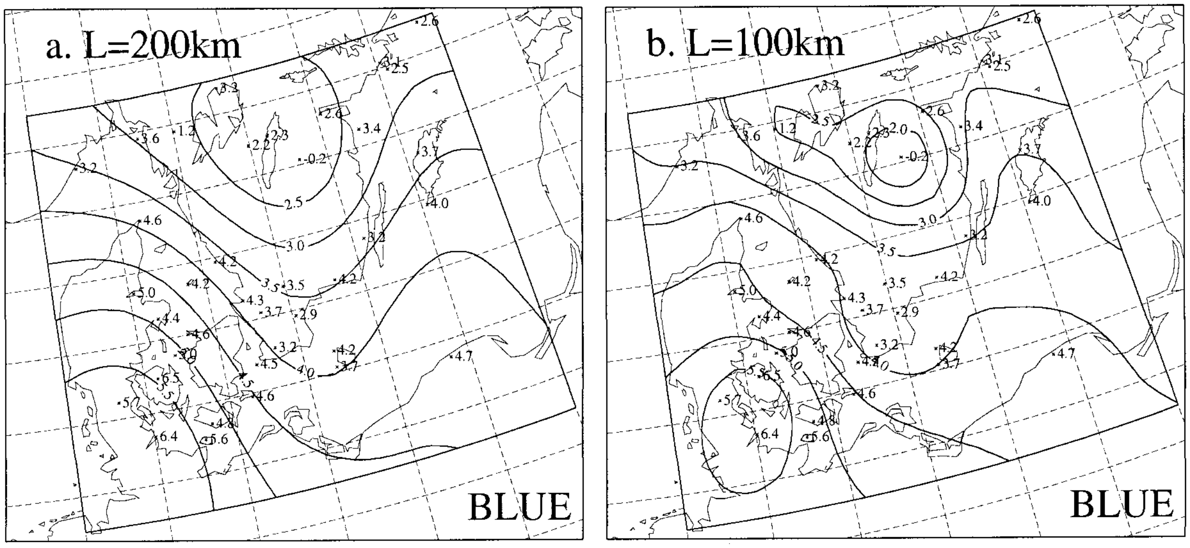

FIG. 3. BLUE analyses using (a) $L = 200$ km and (b) $L = 100$ km.

Using these numbers, a 20th-order filter and a 10th-order filter are chosen in the VAF experiments.

Due to the Lanczos window, the effective scale of the Gaussian filter is reduced. To compensate this scale reduction, $L_f = 260$ km is chosen for the 20th-order filter and $L_f = 128$ km is chosen for the 10th-order filter. (The choice of L_f will be discussed later.) The Gaussian filter coefficients, calculated using (9), as functions of distance are also plotted in Fig. 2, where g260_20 is for a 20th-order filter and g128_10 is for a 10th-order filter. As can be seen from the figure, these two filters closely resemble the two covariance functions.

c. The BLUE solution

The BLUE analyses are shown in Fig. 3. They are obtained by using $L = 200$ km and $L = 100$ km, respectively. As has been discussed in section 2d, the BLUE solution is the final solution of the variational schemes assuming the observation operator is linear and

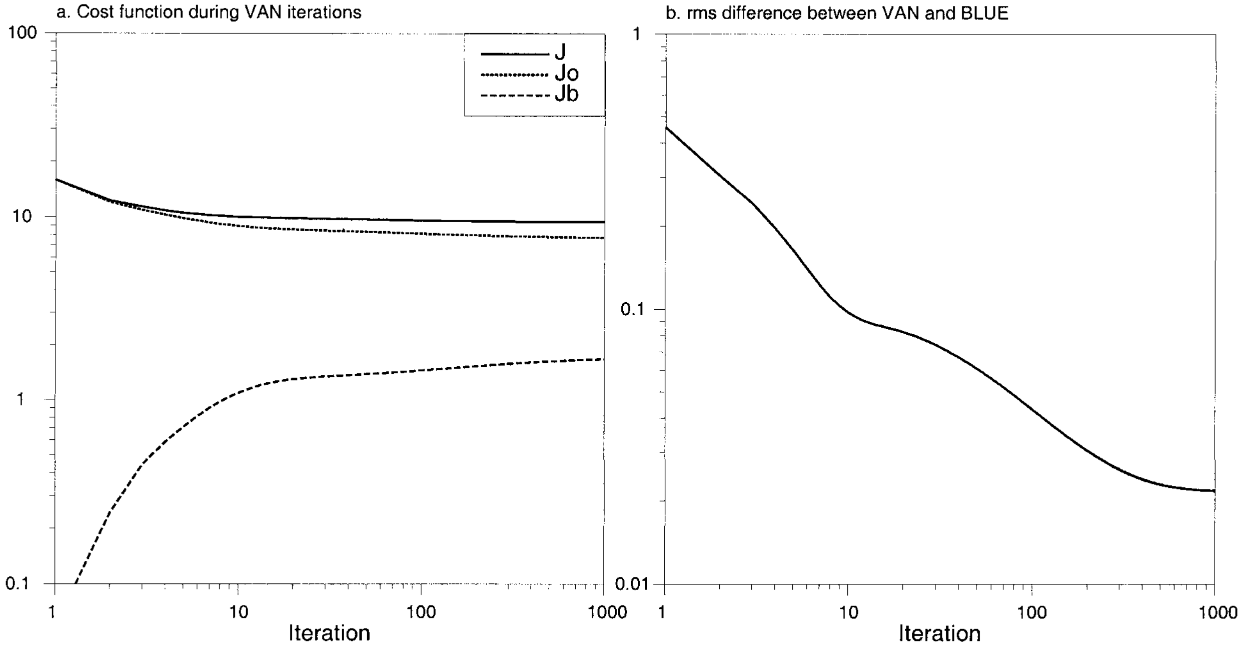

FIG. 4. (a) The cost functions J_o, J_b, and J at each iteration. (b) The rms difference in °C between VAN solution at each iteration and the BLUE solution. Logarithmic scales are used.

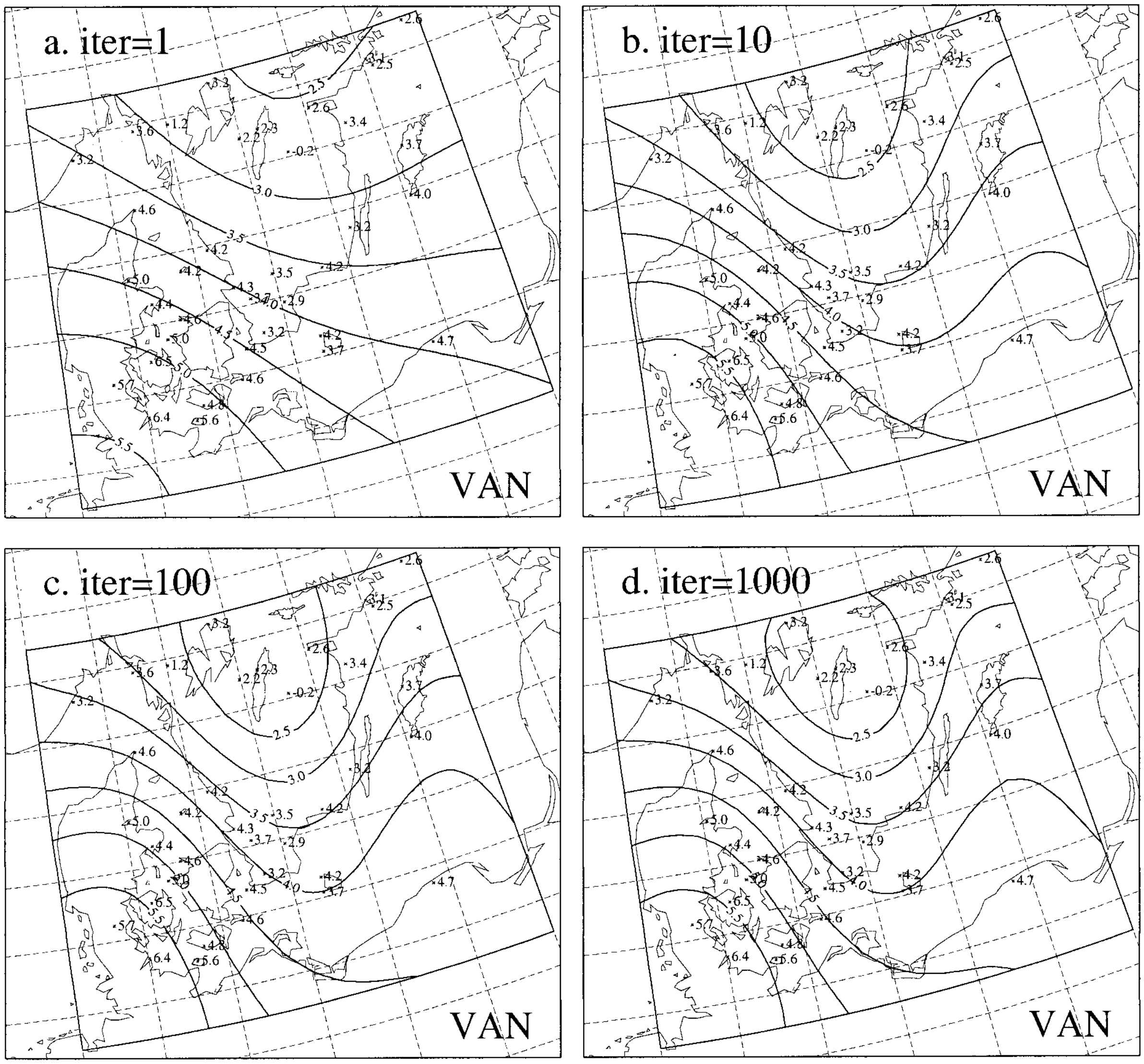

FIG. 5. VAN analyses obtained after (a) 1, (b) 10, (c) 100, and (d) 1000 iterations.

the inversion of ($\mathbf{HBH}^{\mathrm{T}} + \mathbf{R}$) is exact. The inversion is computed using lower triangular and upper triangular (LU) decomposition and routines from Press et al. (1989). This solution is used as a reference to compare the results in the following subsections.

d. Minimization

To solve the variational problems discussed in section 2 an algorithm for minimizing the cost function is needed. In this study, for illustration purposes, the steepest descent minimization algorithm is used. There is no general rule for the choice of the minimization step size, α, used in (4). In the experiments described in this paper, $\alpha = 0.002$ is chosen, which is approximately the largest value with which the steepest descent minimization method converges for all the experiments.

e. VAN solutions

In the experiment using the VAN scheme, the parameters are set to be the same as those used to obtain the BLUE solution with $L = 200$ km. The cost functions J_o, J_b, and J at each iteration (up to 1000) are plotted in Fig. 4a. Note that a logarithmic scale is used.

The choice of the analysis grid, the number of observations, the observation, and back ground error covariances leads to $J_o/J_b \sim 5$; that is, the difference between the cost functions of the analysis and observations is about five times larger than that between analysis and background. With different parameters this ratio can change. As can be seen from Fig. 4a, the largest changes in the cost function occur during the

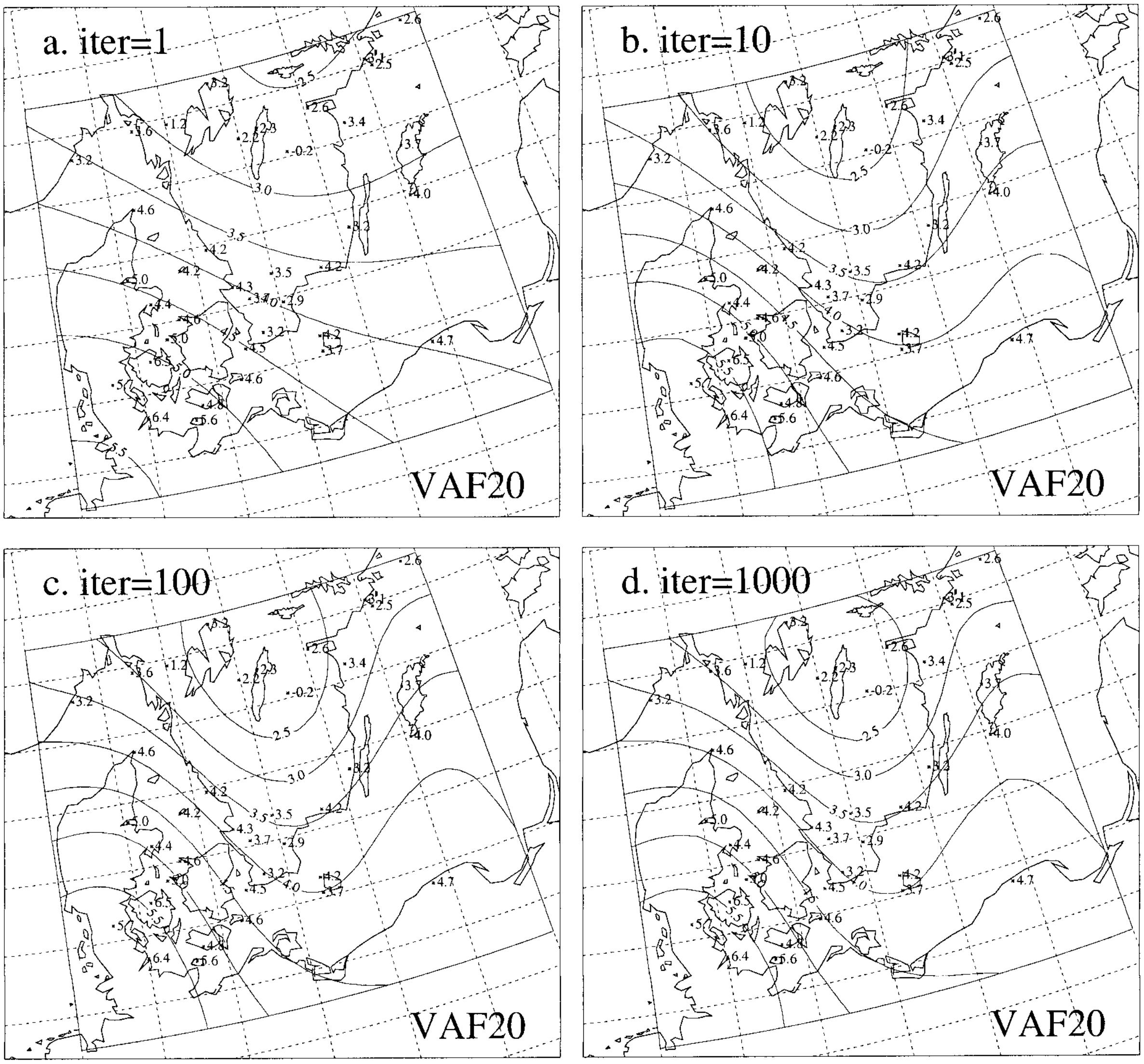

FIG. 6. VAF analyses using a 20th-order Gaussian filter after (a) 1, (b) 10, (c) 100, and (d) 1000 iterations.

first iterations. Even 10 iterations are enough for practical purposes.

Since the BLUE solution could be considered as the final solution for VAN iterations, the rms difference between the VAN solution at each iteration and the BLUE solution is plotted in Fig. 4b. Using this measure, it is shown that the convergence to the final solution could take more time. However, the rms difference is less than 0.1°C after 10 iterations.

The VAN analyses after 1, 10, 100, and 1000 iterations are shown in Fig. 5. The VAN analysis obtained after one iteration is already quite different from the background and has many features of the BLUE solution. The VAN analyses after 10, 100, and 1000 iterations could be considered the same using any subjective judgment, although minor differences, for instance close to the northern boundary, do exist.

f. VAF solutions

Using the Gaussian filter (9) with different orders, a series of experiments using the VAF scheme are conducted. First, a 40th-order filter without the Lanczos window is used as a check of the coding, since the VAF scheme is equivalent to the VAN scheme when the filter parameter $(0.5I + 1) \times (0.5J + 1)$ is equal to the number of grid points of the analysis domain $N_\lambda \times N_\phi$, which are all 21 × 21. The VAF results (not shown) are identical to the VAN results as expected.

Reducing the filter order by a factor of 2 ($I = J = 20$), the solutions (Fig. 6) are practically the same as the VAN solutions (Fig. 5). The solutions after 10 iterations are actually very close to the BLUE solution using $L = 200$ km (Fig. 3a).

A further halving of the filter order ($I = J = 10$)

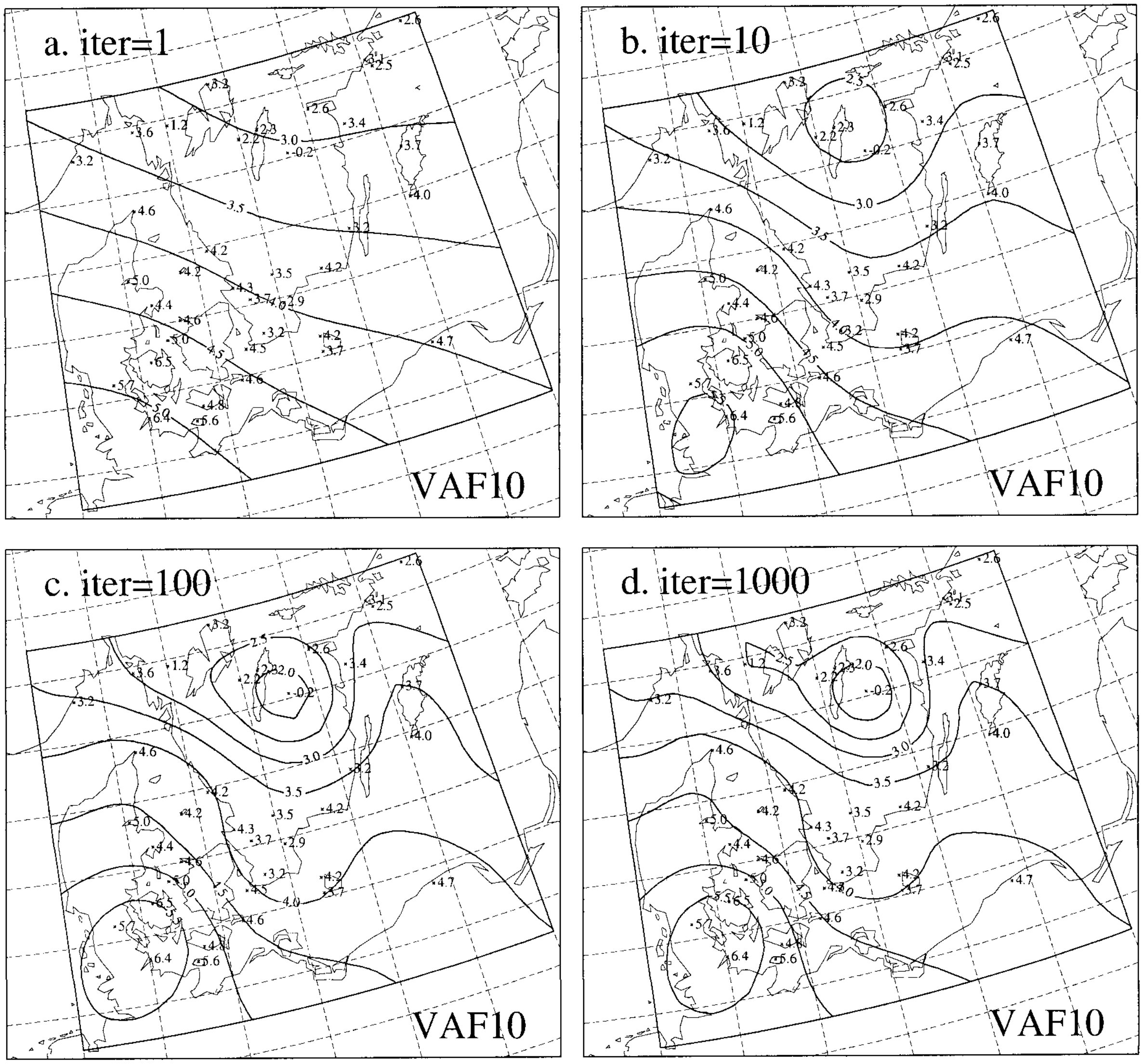

FIG. 7. VAF analyses using a 10th-order Gaussian filter after (a) 1, (b) 10, (c) 100, and (d) 1000 iterations.

leads to a VAF solution with a smaller spatial scale (Fig. 7). As expected from the characteristics of the filter, the solutions converge to the BLUE solution using $L = 100$ km (Fig. 3b).

g. Selection of L_f for VAF

Ideally the background error covariance matrix, $\mathbf{B}$, is computed from statistics and there is nothing left to be tuned. In practice, some kind of analytical function, for example, the Gaussian function (8), is used to model b_{ij} and, in this case, the length scale L is computed from statistics. In other words, L should not be a free parameter to be adjusted.

As discussed in sections 2c and 3b, the length scale L_f used in VAF schemes is not necessarily the same as L due to the filter truncation and window selection. In most cases, L_f should be longer than L.

Using the BLUE solution as the ''true solution'' and the VAN solution as a reference solution, numerical experiments have been performed varying L_f. The results are summarized in Fig. 8, where the rms difference between VAF and BLUE solutions are shown as a function of iterations for the selected experiments. Note that logarithmic scales are used. There are two groups of experiments, one for VAN using $L = 200$ km and the 20th-order filter (Fig. 8a), and the other for VAN using $L = 100$ km and the 10th-order filter (Fig. 8b).

The selection of $L_f = 260$ km for $L = 200$ km and $L_f = 128$ km for $L = 100$ km was based on the final solution of each experiment, which has the minimum in the rms difference. As can be seen from the figure,

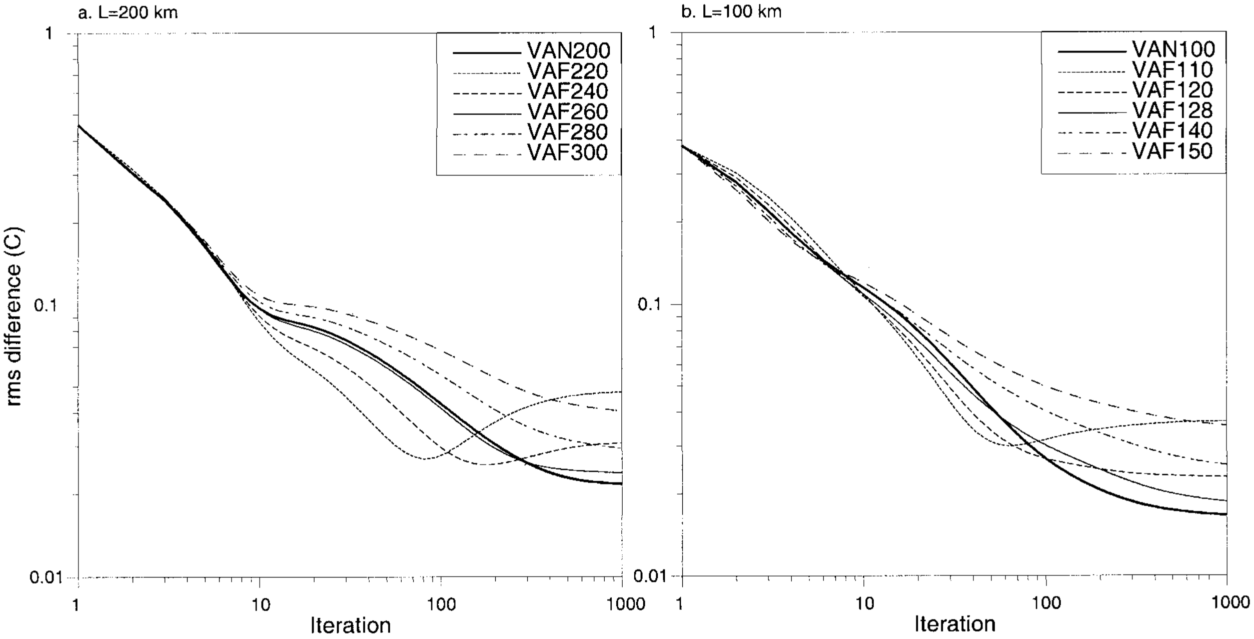

FIG. 8. The rms difference in °C between VAN/VAF and BLUE solutions at each iteration for (a) VAF200 ($L = 200$ km) and VAF schemes with different L_f (from 220 km to 300 km) and (b) VAF100 ($L = 100$ km) and VAF schemes with different L_f (from 110 km to 150 km). Logarithmic scales are used.

the VAN solution is best at the end of iterations. This should be expected, since VAF is just an approximation to VAN.

A closer inspection of Fig. 8 also reveals a few interesting aspects of the VAF solutions. First of all, within the first 10 iterations the VAF solutions are not sensitive to L_f for a large range of L_f. This means that in practical applications the precise value of L_f may not be crucial. Second, there is an iteration range between 10 and 100 in which VAF solutions could be better than the VAN solutions. In practical implementations, there is often a predetermined maximum iteration number that could be in this range. In that case, it may be possible to improve the analysis by taking a L_f that is somewhat shorter than that selected based on the final solutions. Third, the distance between some VAF solutions and the "true solution" decreases through the early iterations and reaches the minimum at, for example, around 100 iterations. Further iterations would lead to an increase in the distance. Finally, it should be stressed that the rms difference between different solutions is plotted on logarithmic scales in Fig. 8 and is much smaller than the assumed errors in both background and observations.

4. A more realistic example

To give a more realistic picture, some of the experiments described in section 3 have been repeated with the following changes:

1) An increase of analysis grid points from 21 × 21 to 101 × 101.
2) An increase of observations from 38 to 1062.

Other parameters, for example, $\Delta\lambda$, $\Delta\phi$, σ_o, and σ_b, remain unchanged except that $\alpha = 0.0005$. The dimension of $\mathbf{B}$ is now 10^8, which makes the VAN scheme impossible to run even on some supercomputers. Therefore only a VAF solution is shown. The dimension of ($\mathbf{HBH}^T + \mathbf{R}$) is now 10^6, which makes the matrix inversion of the BLUE scheme nontrivial to handle.

In this example, 2-m temperature observations at 0000 UTC 5 May 1999 within 35°–65°N and 10°W–50°E are selected from the Danish Meteorological Institute operational database. The locations and values of the observations are given by Fig. 9. The full analysis domain is shown in the figure. As in the previous section, the background field has the reasonable, large-scale structure of the observations (Fig. 9a).

Both the BLUE solution (with $L = 200$ km; Fig. 9b) and the VAF solution (with $L_f = 260$ km at the 1000th iteration; Fig. 9c) add smaller-scale information from observations to the background. Comparing the two analyses subjectively, it is obvious that the VAF solution is a good approximation of the BLUE solution. To give a quantitative comparison, the rms difference between the VAF and BLUE solutions as a function of iterations is plotted in Fig. 10. Based on this figure and the results from the previous section, it could also be argued that 10 iterations may be enough for the VAF scheme to

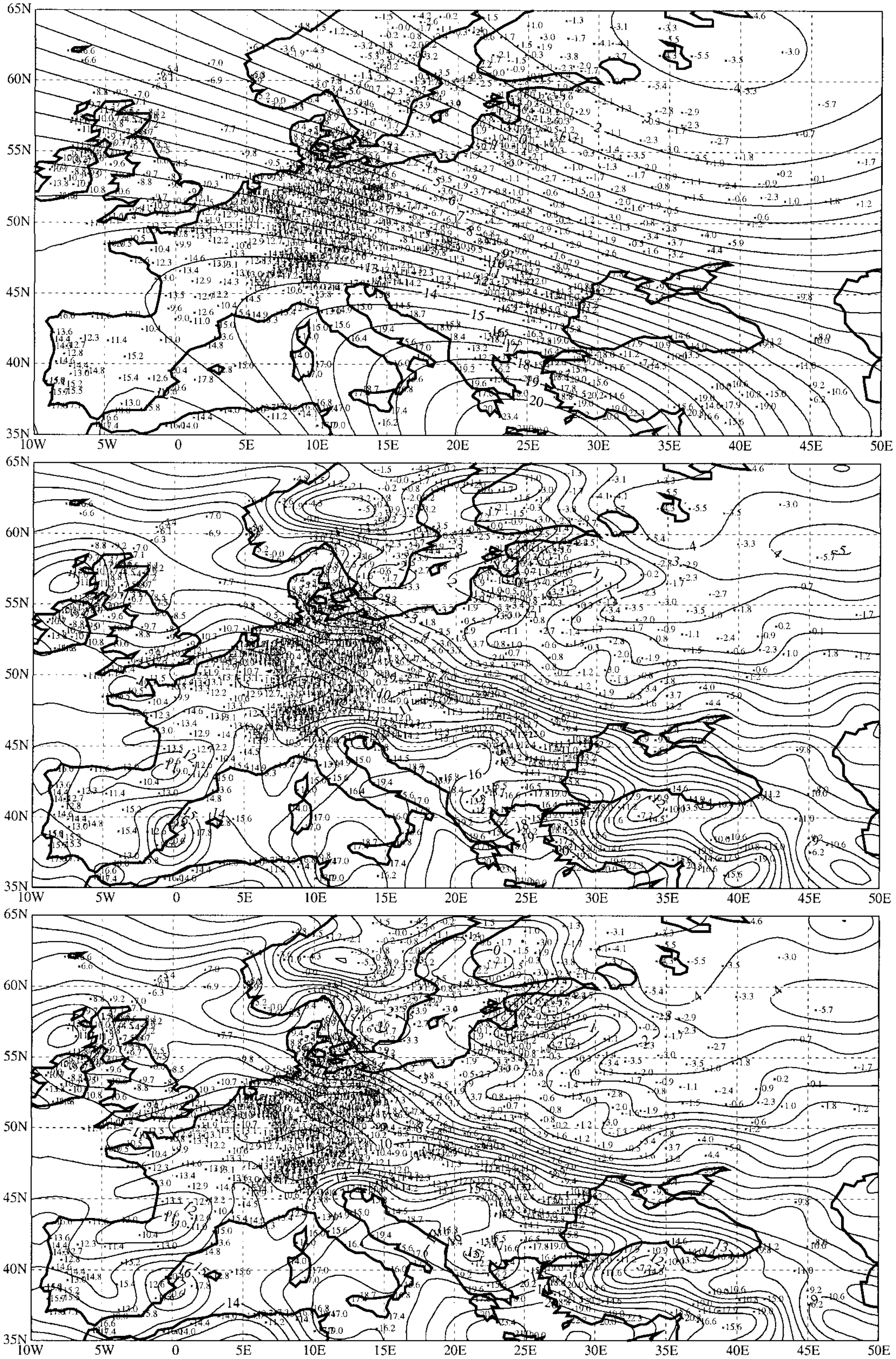

FIG. 9. The 2-m temperature observations (locations are indicated by stars and values in °C) at 0000 UTC 5 May 1999 and the (a) background field, (b) BLUE analysis, and (c) VAF analysis.

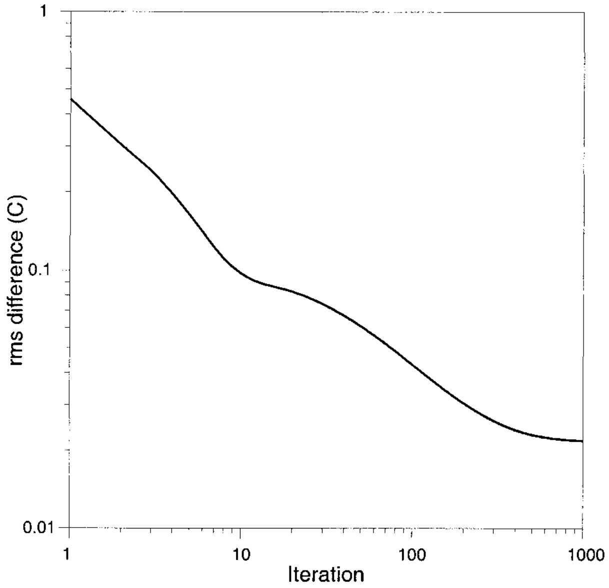

FIG. 10. The rms difference in °C between VAF and BLUE solutions at each iteration. Logarithmic scales are used.

produce an analysis of approximately the same quality as the BLUE solution.

5. Conclusions

In this paper, we have addressed two issues related to the variational data assimilation, VAR, as formulated by Lorenc (1986). The first is how to avoid the inversion of the large dimensioned background covariance matrix **B**. The second is how to avoid explicit storage and computation of **B**.

Following Lorenc (1988) and Derber and Rosati (1989), we have shown that a simple transform of the control variable leads to a modified scheme, named VAN—variational analysis with no inversion of the background error covariance matrix. The remaining storage and computation difficulties in VAN are mainly due to the presence of the matrix itself. Although a full-sized implementation of the VAN scheme is still difficult, the VAN scheme is considered as a step forward from VAR. The problems related to the inversion of a very large and often ill-conditioned matrix are avoided.

A limitation of the VAN scheme is the lack of preconditioning. If some preconditioning, for example, using the inverse of the Hessian matrix, is needed, the inversion of the background error covariance matrix may enter the preconditioning and approximations would be needed to avoid inversions.

Based on the VAN scheme, we have formulated a new scheme, named VAF—variational analysis using a filter, which uses a spatial filter to model the background error covariance matrix and does not require the matrix explicitly. Using a truncated filter the spatial scale of the analysis is changed slightly. A redefinition of the length scale in the filter is necessary to adjust the effective scale of the analysis. The storage and computation required by the VAF scheme could be orders of magnitude smaller than the VAR and VAN schemes. A full-sized implementation of the VAF scheme is possible.

To illustrate how these schemes work, a small-sized and a more realistic two-dimensional univariate analysis problem are considered using real 2-m temperature observations. As the observation operator involved is linear, the best linear unbiased estimates are used as reference. It is shown that both VAN and VAF work satisfactorily.

In this paper, a Gaussian function is used to model **B**, which leads naturally to a Gaussian filter in VAF. In realistic applications, the form of **B** should be tuned statistically using both background and observations. Other forms of **B**, for example, a series of Bessel functions, may also be constructed and used to determine the filter coefficients for VAF. However, without a meaningful verification package it is difficult to compare different **B** formulations. Further studies of this are needed.

In the construction of the spatial filter, a Lanczos window is introduced to reduce the truncation-related Gibbs oscillations. During the development of the VAF scheme, many problems caused by noise have occurred when the truncation was severe and the window was inactive. The Lanczos window is our first attempt and seems to be quite effective in removing noise. Although other windows may be used, a proper comparison method needs to be established to select a better window.

One of the advantages of using VAR, VAN, or VAF is the ability of assimilating indirect data. We plan to use the VAF scheme to include the ground-based Global Positioning System data in the type of analyses shown here. A more sophisticated minimization algorithm and a better preconditioning algorithm, for example as discussed by Dévényi and Benjamin (1998), will have high priority in the future development. A nonuniform grid version of VAF needs to be derived, as shown by Lorenc (1992), for more realistic implementations. A full-sized three-dimensional multivariant problem should be considered. In order to make efficient use of asynoptic data and to obtain a coherent four-dimensional analysis, a time-dependence and a numerical weather prediction model should also be included by the VAN and VAF schemes. These will be research subjects in the future.

Acknowledgments. The author would like to thank Dezsö Dévényi, Nils Gustafsson, Erland Källén, Leif Laursen, Andrew Lorenc, Peter Lynch, and Henrik Vedel for comments on an earlier version of the manuscript. Comments from anonymous reviewers are also appreciated. This work has partly been supported by the Danish Research Councils through Grant 9702667.

REFERENCES

Barnes, S., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.,* **3,** 395–409.

Bergthorsson, P., and B. Döös, 1955: Numerical weather map analysis. *Tellus,* **7,** 329–240.

Cohn, S. E., A. Da Silva, J. Guo, M. Sienkiewicz, and D. Lamich, 1998: Assessing the effects of data selection with the DAO Physical-space Statistical Analysis System. *Mon. Wea. Rev.,* **126,** 2913–2926.

Courtier, P., 1997: Variational methods. *J. Meteor. Soc. Japan,* **75,** 211–218.

——, and Coauthors, 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. *Quart. J. Roy. Meteor. Soc.,* **124,** 1783–1808.

Cressman, G., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **87,** 367–374.

Daley, R., 1991: *Atmospheric Data Assimilation.* Cambridge University Press, 457 pp.

Derber, J., and A. Rosatti, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.,* **19,** 1333–1347.

Dévényi, D., and S. G. Benjamin, 1998: Application of a three-dimensional variational analysis in RUC-2. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 37–41.

Eliassen, A., 1954: Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. Rep. 5, Videnskaps-Akademiet, Institut for Vaer och Klimaforskning (Norwegian Academy of Sciences, Institute of Weather and Climate Research), Oslo, Norway, 12 pp.

Gandin, L., 1963: *Objective Analysis of Meteorological Fields.* Gidromet, Leningrad, 285 pp. English translation, Israel Program for Scientific Translations, 1965.

Gustafsson, N., and Coauthors, 1999: Three-dimensional variational data assimilation for a high resolution limited area model (HIRLAM). HIRLAM Tech. Rep., 40, 74 pp. [Available from Met Éireann, Glasnevin Hill, Dublin 9, Ireland.]

Hayden, C. M., and R. J. Purser, 1995: Recursive filter objective analysis of meteorological fields: Applications to NESDIS operational processing. *J. Appl. Meteor.,* **34,** 3–15.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. *J. Meteor. Soc. Japan,* **75,** 181–189.

Lönnberg, P., and D. Shaw, Eds., 1987: ECMWF data assimilation—Scientific documentation. ECMWF Research Manual 1. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

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

——, 1988. Optimal nonlinear objective analysis. *Quart. J. Roy. Meteor. Soc.,* **114,** 205–240.

——, 1992: Interactive analysis using covariance functions and filters. *Quart. J. Roy. Meteor. Soc.,* **118,** 569–591.

——, 1997: Development of an operational variational assimilation scheme. *J. Meteor. Soc. Japan,* **75,** 339–346.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's global spectral statistical interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1989: *Numerical Recipes. The Art of Scientific Computing (FORTRAN Version).* Cambridge University Press, 702 pp.