A Preliminary Study on Dimension-Reduction Algorithm for Variational Methods in Three Dimensions

Xuan Chen*[1]

[1]*Numerical Analysis Research Center, GZ 510510, China*

ABSTRACT

Three Dimensional Variational data assimilation or analysis (3DVAR) is one of most classical methods for providing the initial values for numerical models. In this method, the dimensions of the background error covariance and the observational error covariance matrices are large. Therefore, it is difficult to get the inverse of the covariance matrices and to reduce the orders of these matrices without information loss. With the use of the Sylvester Equation, on the basis of a new linear regression, a new cost function for 3DVAR was given. For the first-guess $m \times n$ field, there is an approximate $1-(m^2+n^2)/(mn \times mn)$ reduction with $m>1$ & $n>1$ by using the cost function. The results of the numerical experiments show that the effect of this algorithm is no worse than that of the old cost function for 3DVAR.

Key words: 3DVAR; data assimilation; cost function; Sylvester equation

1. Introduction

Variational data assimilation is widely used all over the world (Chen et al., 2015; Kalnay, 2003; Kristian et al., 2009; Warner, 2011) in order to provide the first guesses (initial values) for atmospheric or oceanic numerical predictions, which is an initial

*Corresponding author: Chen Xuan (E-mail: chzffx@qq.com or chenxuan@mail.iap.ac.cn)

value problem. In these methods, the background error covariance and observational error covariance matrices have large dimensions (Kalnay, 2003; Warner, 2011). Reducing the order of the covariance matrices is still an important problem.

It is well known that the structure or estimation of the covariance matrices in the cost function (Bannister, 2008a; Bannister, 2008b; Federico, 2013; Liu et al., 2010; Parrish et al., 1992) is the key to solving the variational analysis problem, but this type of matrix is nearly ill-conditioned in the large dimension case. In order to get a better estimation of these matrices, many algorithms (Bannister, 2008a; Bannister, 2008b; Cohn et al., 1996; Farrell et al., 2010) for these large matrices have been developed. Some of them are applied by many meteorological centers, such as the NMC (National Meteorological Center) method (Parrish et al., 1992). Naturally, reduction-order algorithms are also developed; proper orthogonal decomposition is one of these methods, and many works on reduction-order methods are based on it (Cao et al., 2007; Fang et al., 2014; Lawless et al., 2008). The others are based on statistical methods, such as empirical orthogonal function (EOF) decomposition (Frolov et al., 2009; Hoteit et al., 2006; Robert et al., 2005; Shen et al., 2014; Zhao et al., 2012) and the work of Kleist et al. (2008). By reducing the computational burden, more or less information of the covariance matrix is neglected. It is obvious that both keeping the covariance matrices' information and reducing the computational burden are important. Although those methods are effective, fewer of them can truly reduce the dimensions of the covariance matrices.

In previous research (Chen et al., 2017; 2019), a more generalized linear regression model was developed. By using this method, Chen et al. (2019) found that the Sylvester Equation (SE) can be applied in data assimilation and that the SE can reduce the computational burden with less information loss. In this paper, the theory of the SE and the use of the SE to solve the Three Dimensional Variational analysis (3DVAR) problem are presented in section 2, some numerical experiments are presented in section 3, and the conclusions and a discussion are provided in section 4.

2. Reduction-order algorithm

2.1 The SE

Eq. (1) in the $m \times n$ unknown matrix $\mathbf{X}$ is called the SE as follows:

$$\mathbf{AX}+\mathbf{XB}+\mathbf{C}=\mathbf{0}, \tag{1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $m \times m$ and $n \times n$ square matrices, respectively, and $\mathbf{C}$ is $m \times n$ (Simoncini, 2013). Eq. (1) is a linear matrix equation and is widely used in controlling systems, numerical analysis and even image processing.

Eq. (1) can be rewritten as the following linear vector equation:

$$(\mathbf{I}_n \otimes \mathbf{A}+\mathbf{B}^{\mathrm{T}} \otimes \mathbf{I}_m)\mathbf{vec}(\mathbf{X})+\mathbf{vec}(\mathbf{C})=\mathbf{0}, \tag{2}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\otimes$ is the Kronecker product notation, $\mathbf{T}$ represents the matrix transposition, and **vec**() stacks the columns of $\mathbf{X}$ into a column vector (Deif et al., 1995; Simoncini, 2013).

2.2 The more generalized linear regression model

Chen et, al. (2017; 2019) proposed a more generalized linear regression model, whose basic features include non-commutative multiplication, equivalence to traditional linear regression as predictors in the model are scalars, analysis, extension, dimension-reduction, and robustness, etc., and applied this model in the hind-forecasting experiment of seasonal precipitation in China, the correlation coefficients of the precipitation from NCEP and from Chen's model are greater than 0.7 in the most areas, and the mean absolute errors are less than 0.8 in the most areas. So, this model is applicable in practice.

For linear regression model, in a simple way, $y=ax+b$, we can estimate the coefficients a and b by least square method. Then it is easy and natural to extent the model to the both vector form: $[y_1, y_2, y_3, \dots, y_m]^{\mathrm{T}}=\mathbf{A}_{m\times n}[x_1, x_2, x_3, \dots, x_n]^{\mathrm{T}}+[b_1, b_2, b_3, \dots, b_m]^{\mathrm{T}}$, which can also be solved by least square method. But for more generalized linear regression model, this expression seems to be limited by its form. For the expression for the matrix-form data $\mathbf{Y}_{m\times p}$ and $\mathbf{X}_{n\times q}$, it is difficult to solving directly by using the least square method, but the most important principle is all the same that the error of the regression model must be smallest.

So, for the matrix-form data $\mathbf{Y}_{m\times p}$ and $\mathbf{X}_{n\times q}$, the simplest regression model is $\mathbf{Y}_{m\times p}=\mathbf{A1}_{m\times n}\mathbf{X}_{n\times q}\mathbf{A2}_{q\times p}+\mathbf{B}_{m\times p}$. How to estimate the coefficients matrixes? Taking the most important principle about the error into consideration, here noting the matrix $\mathbf{Y}_{m\times p}$ in the model as $\hat{\mathbf{Y}}_{m\times p}$, as taking $\mathbf{Y}_{m\times p}$ as observations. The error of the model can be written as $\sum(y_{ij}-\hat{y}_{ij})^2$, which is the sum of two matrixes' trace of the expression: $[\mathbf{Y}_{m\times p}-\hat{\mathbf{Y}}_{m\times p}][\mathbf{Y}_{m\times p}-\hat{\mathbf{Y}}_{m\times p}]^{\mathrm{T}}$ and $[\mathbf{Y}_{m\times p}-\hat{\mathbf{Y}}_{m\times p}]^{\mathrm{T}}[\mathbf{Y}_{m\times p}-\hat{\mathbf{Y}}_{m\times p}]$. With using matrix differential

rule, we can get the expression of coefficients matrixes but nonlinear. By using the Sylvester Equation, it seems easier for solving: at first, giving the estimation of the coefficients matrixes $\mathbf{A1}$ by the expression $\hat{\mathbf{Y}}\mathbf{1}_{m\times\max(p,q)}=\mathbf{A1}_{m\times n}\mathbf{X}_{n\times\max(p,q)}+\mathbf{B1}_{m\times\max(p,q)}$, in which the elements in the part from min (p,q) to max(p,q) of matrix $\mathbf{X}$ or $\mathbf{Y}$ can be filled with random numbers. For example, p=min(p,q) and q=max(p,q), $[\hat{\mathbf{Y}}\mathbf{1}_{m\times p},\mathbf{R}_{m\times(q-p)}]=\mathbf{A1}_{m\times n}\mathbf{X}_{n\times q}+\mathbf{B1}_{m\times q}$, where matrix $\mathbf{R}_{m\times(q-p)}$ is filled with random numbers. By the least square method and the Sylvester Equation, we can estimate $\mathbf{A1}$ and $\mathbf{B1}$. Then by the form $\hat{\mathbf{Y}}_{m\times p}=[\hat{\mathbf{Y}}\mathbf{1}_{m\times p},\mathbf{R}_{m\times(q-p)}]\mathbf{A2}_{q\times p}+\mathbf{B2}_{m\times p}$, and by the same method, we can estimate the coefficients $\mathbf{A2}$ and $\mathbf{B2}$. At last, deleting the random part, the expression of more generalized linear regression model is $\mathbf{Y}_{m\times p}=\mathbf{A1}_{m\times n}\mathbf{X}_{n\times q}\mathbf{A2}_{q\times p}+\mathbf{B}_{m\times p}$, where $\mathbf{B}_{m\times p}=\mathbf{B1}_{m\times q}\mathbf{A2}_{q\times p}+\mathbf{B2}_{m\times p}$.

2.3 3DVAR

In 3DVAR, the scale cost function $J(\mathbf{x})$ is defined as follows:

$$J(\mathbf{x})=(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)+(\mathbf{y}-\mathbf{H}[\mathbf{x}])^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}[\mathbf{x}]), \quad (3)$$

where $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}$ is the observational error covariance matrix, $\mathbf{x}$ is the unknown vector, $\mathbf{x}_b$ is the background or the first guess vector, $\mathbf{y}$ is the observation vector and $\mathbf{H}[]$ is the observation operator and assumed to be linear. The cost function can be defined as the distance between the analysis and the background, weighted by the inverse of the background error covariance, plus the distance to the observations, weighted by the inverse of the observations error covariance (Kalnay, 2003). And, Eq. (1) can be written as:

$$J(\mathbf{x})=\mathrm{Trace}[(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)]+\mathrm{Trace}[(\mathbf{y}-\mathbf{H}[\mathbf{x}])^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}[\mathbf{x}])],$$

where Trace[] is the trace of the matrix in square brackets. Eq. (1) is minimized in order to obtain the analysis by solving the following equation:

$$\nabla J(\mathbf{x})=2\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)+2\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}-\mathbf{x}_b)-2\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}[\mathbf{x}_b])\equiv\mathbf{0}, \quad (4)$$

after solving Eq. (4), the following equation is used to obtain the analysis vector:

$$\mathbf{x}_a=\mathbf{x}_b+\mathbf{\Delta}, \quad (5)$$

where $\mathbf{\Delta}=(\mathbf{B}^{-1}+\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}[\mathbf{x}_b])$, one form of the solution of equation (4).

Taking the one point situation into consideration, the cost function J(x) can be written as:

$$J(x)=(x-x_b)b^{-1}(x-x_b)+(y-x)r^{-1}(y-x)$$

taking the same views that the cost function is a quadratic function of the analysis increments $(x-x_b)$, then the gradient of the cost function J with respect to $(x-x_b)$ or equation (4) can be written as:

$$\nabla J(x)=2b^{-1}(x-x_b)+2r^{-1}(y-x_b)\equiv 0$$

then the analysis can be written as:

$$x_a=(b^{-1}+r^{-1})^{-1}b^{-1}x_b+(b^{-1}+r^{-1})^{-1}r^{-1}y$$

which is minimum variance estimator and for the error (ε_a, ε_b, ε_o (for y)) expression,

$$\varepsilon_a=(b^{-1}+r^{-1})^{-1}b^{-1}\varepsilon_b+(b^{-1}+r^{-1})^{-1}r^{-1}\varepsilon_o$$

For the vector form, we can also get the form by set $(b^{-1}+r^{-1})^{-1}b^{-1}$ and $(b^{-1}+r^{-1})^{-1}r^{-1}$ of each elements in the verctor $\mathbf{x}_b$ and $\mathbf{y}$ as element of the weight diagonal matrixes $\mathbf{D}_1$ and $\mathbf{D}_2$ in $\mathbf{x}=\mathbf{D}_1\mathbf{x}_b+\mathbf{D}_2\mathbf{H}^{-1}(\mathbf{y})$ respected to the elements in the vector, here $\mathbf{H}^{-1}(\mathbf{y})$ just means we can get the point information, and with any other practical meaning. The errors of the analysis ($\boldsymbol{\varepsilon}$) can be considered as the sum of the

errors of background ($\boldsymbol{\varepsilon}_b$) and observation ($\boldsymbol{\varepsilon}_o$): $\boldsymbol{\varepsilon}=\mathbf{D}_1\boldsymbol{\varepsilon}_b^{\mathrm{T}}+\mathbf{D}_2\mathbf{H}^{-1}\left(\boldsymbol{\varepsilon}_o\right)$. Then, what to do is just giving the estimation of the minimum variance. Same to the point situation, and just considering all errors' distribution is same, we can give the coefficients a_1 and a_2 in $\mathrm{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)=\mathrm{E}\left(a_1\boldsymbol{\varepsilon}_b^{\mathrm{T}}\boldsymbol{\varepsilon}+a_2\mathbf{H}^{-1}\left(\boldsymbol{\varepsilon}_o\right)^{\mathrm{T}}\mathbf{H}^{-1}\left(\boldsymbol{\varepsilon}_o\right)\right)$, where a_1+a_2=1 and a_1 and a_2 are constants related to the variances of the errors with the assumption that the parameters of statistic characters in each grid and observation station are same..

All of this is easy, but difficult for the matrix form. So, by using the more generalized linear regression model, just give the forms with the more generalized linear regression model. On the basis of matrix theory and a more generalized linear regression model (Chen et al., 2017; 2019), the analysis $\mathbf{X}$ ($\mathbf{x}=\mathbf{vec}(\mathbf{X})$, $\mathbf{X}$ is the analysis field and $m\times n$) can be written as:

$$\mathbf{X}=\mathbf{E}_1\mathbf{X}_b\mathbf{E}_2+\mathbf{E}_3\mathbf{H}^{-1}(\mathbf{Y})\mathbf{E}_4.$$

where, $\mathbf{E}_i$ (i=1, 2, 3, 4) is the regression coefficient matrices, $\mathbf{H}[]$ is the observation operator and assumed to be linear, $\mathbf{H}^{-1}$ is the generalized inverse to the $\mathbf{H}$. Because $\mathbf{X}$, $\mathbf{X}_b$ and $\mathbf{H}^{-1}(\mathbf{Y})$ are the different forms of the same field, the regression coefficient matrices are the identity matrices weighted a constant number and the $\mathbf{E}_1\mathbf{E}_2+\mathbf{E}_3\mathbf{E}_4=\mathbf{1}$. The errors of the analysis ($\boldsymbol{\varepsilon}$) can be considered as the sum of the errors of background ($\boldsymbol{\varepsilon}_b$) and observation ($\boldsymbol{\varepsilon}_o$):

$$\boldsymbol{\varepsilon}=\mathbf{E}_1\boldsymbol{\varepsilon}_b\mathbf{E}_2+\mathbf{E}_3\mathbf{H}^{-1}(\boldsymbol{\varepsilon}_o)\mathbf{E}_4.$$

The distance between the analysis and the truth field can be also written as $\mathrm{Trace}(\boldsymbol{\varepsilon}\,\mathbf{W}_1^{-1}\boldsymbol{\varepsilon}^{\mathrm{T}})+\mathrm{Trace}(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{W}_2^{-1}\boldsymbol{\varepsilon})$, where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the expectations of $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}$. Taking the expectation of $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ as an example:

$$\mathrm{E}(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon})=\mathrm{E}[(\mathbf{E}_1\boldsymbol{\varepsilon}_b\mathbf{E}_2+\mathbf{E}_3\mathbf{H}^{-1}(\boldsymbol{\varepsilon}_o)\mathbf{E}_4)^{\mathrm{T}}(\mathbf{E}_1\boldsymbol{\varepsilon}_b\mathbf{E}_2+\mathbf{E}_3\mathbf{H}^{-1}(\boldsymbol{\varepsilon}_o)\mathbf{E}_4)],$$

on the basis of the assumptions above, the expectation can be written as:

$$\mathrm{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)=\mathrm{E}\left(a_1^2\boldsymbol{\varepsilon}_b^{\mathrm{T}}\boldsymbol{\varepsilon}_b+a_2^2\mathbf{H}^{-1}\left(\boldsymbol{\varepsilon}_o\right)^{\mathrm{T}}\mathbf{H}^{-1}\left(\boldsymbol{\varepsilon}_o\right)\right).$$

where a_1 and a_2 are constants related to the variances of the errors; in addition, a_1+a_2=1, and both a_1 and a_2 are greater than zero. So, it is natural to rewrite the cost function as the following equation:

$$J(\mathbf{X})=J_1(\mathbf{X})+J_2(\mathbf{X})+J_3(\mathbf{X})+J_4(\mathbf{X}), \tag{6}$$

where

$$\begin{cases} J_1\left(\mathbf{X}\right)=\mathrm{Trace}\left[\left(\mathbf{X}-\mathbf{X}_b\right)^{\mathrm{T}}\mathbf{B}_1^{-1}\left(\mathbf{X}-\mathbf{X}_b\right)\right] \\ J_2\left(\mathbf{X}\right)=\mathrm{Trace}\left[\left(\mathbf{X}-\mathbf{X}_b\right)\mathbf{B}_2^{-1}\left(\mathbf{X}-\mathbf{X}_b\right)^{\mathrm{T}}\right] \\ J_3\left(\mathbf{X}\right)=\mathrm{Trace}\left[\left(\mathbf{Y}-\mathbf{H}\left[\mathbf{X}\right]\right)^{\mathrm{T}}\mathbf{R}_1^{-1}\left(\mathbf{Y}-\mathbf{H}\left[\mathbf{X}\right]\right)\right] \\ J_4\left(\mathbf{X}\right)=\mathrm{Trace}\left[\left(\mathbf{Y}-\mathbf{H}\left[\mathbf{X}\right]\right)\mathbf{R}_2^{-1}\left(\mathbf{Y}-\mathbf{H}\left[\mathbf{X}\right]\right)^{\mathrm{T}}\right] \end{cases}, \tag{7}$$

where $\mathbf{B}_1$ (with a dimension of $m\times m$) and $\mathbf{B}_2$ (with a dimension of $n\times n$) are the background error covariance matrices, $\mathbf{R}_1$ (with a dimension of $m\times m$) and $\mathbf{R}_2$ (with a dimension of $n\times n$) are the observational error covariance matrices, $\mathbf{Y}$ is the observation and with a dimension of $m\times n$, here, we just fill the observation matrix $\mathbf{Y}$ with all observation data along the diagonal of the matrix which ensure the matrix is a nonsingular matrix, and the others in the matrix just filled with zero. For the vertical levels, just put them together as [$\mathbf{X}_1$;$\mathbf{X}_2$;$\mathbf{X}_3$;....;$\mathbf{X}_n$], then use equation (6) to get the analysis. Therefore, the analysis can be conducted by minimizing Eq. (6). But, it should be noted that the cost function (Eqs.(6) and (7)) is not exactly the same as the regular cost function (Eq. (3)).

2.4 Discussion in the Ideal Case

In the ideal case, the observational stations are located in each computational grid point; in another word, for the ideal case, in each grid, there are two values: the back ground value and the observation value. On the basis of matrix theory and a more generalized linear regression model, $\mathbf{H}[\mathbf{X}]$ can be written as $\mathbf{H}[\mathbf{X}]\equiv\mathbf{AXD}$, where $\mathbf{A}$ and $\mathbf{D}$ are the regression coefficient matrices (in the more generalized linear regression, if there is linear correlation between the two anomalies $\mathbf{X}$ and $\mathbf{Y}$, the relation can be written as $\mathbf{Y}=\mathbf{AXD}$, Chen et al., 2017; 2019).

Differentiating $J_1(\mathbf{X})$ in Eq. (7) with respect to $\mathbf{X}$ gives the following:

$$\frac{\partial J_1(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{B}_1^{-1}\mathbf{X} + \mathbf{B}_1^{-1\mathrm{T}}\mathbf{X} - \mathbf{B}_1^{-1}\mathbf{X}_b - \mathbf{B}_1^{-1\mathrm{T}}\mathbf{X}_b, \tag{8}$$

Similar to other terms, we substitute these equations into the following equation:

$$\frac{\partial J(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}}\left[J_1(\mathbf{X}) + J_2(\mathbf{X}) + J_3(\mathbf{X}) + J_4(\mathbf{X})\right], \tag{9}$$

with

$$\mathbf{\Delta}=\mathbf{X}-\mathbf{X}_b. \tag{10}$$

Then, Eq. (9) can be rewritten as follows:

$$\frac{\partial J(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{B}_{1S}\mathbf{\Delta} + \mathbf{A}^{\mathrm{T}}\mathbf{R}_{1S}\mathbf{A}\mathbf{\Delta}\mathbf{D}\mathbf{D}^{\mathrm{T}} + \mathbf{\Delta}\mathbf{B}_{2S} + \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{\Delta}\mathbf{D}\mathbf{R}_{2S}\mathbf{D}^{\mathrm{T}} - \mathbf{F}_1 - \mathbf{F}_2 \equiv \mathbf{0}, \tag{11}$$

The other terms in Eq. (11) are as follows:

$$\begin{cases} \mathbf{B}_{1S} = \mathbf{B}_1^{-1} + \mathbf{B}_1^{-1\mathrm{T}} \\ \mathbf{R}_{1S} = \mathbf{R}_1^{-1} + \mathbf{R}_1^{-1\mathrm{T}} \\ \mathbf{F}_1 = \mathbf{B}_{1S}\mathbf{X}_b + \mathbf{A}^{\mathrm{T}}\left(\mathbf{R}_1^{-1} + \mathbf{R}_1^{-1\mathrm{T}}\right)\mathbf{G}\mathbf{D}^{\mathrm{T}} + \mathbf{A}^{\mathrm{T}}\mathbf{R}_{1S}\mathbf{A}\mathbf{X}_b\mathbf{D}\mathbf{D}^{\mathrm{T}} \end{cases}, \tag{12}$$

and

$$\begin{cases} \mathbf{B}_{2S} = \mathbf{B}_2^{-1} + \mathbf{B}_2^{-1\mathrm{T}} \\ \mathbf{R}_{2S} = \mathbf{R}_2^{-1} + \mathbf{R}_2^{-1\mathrm{T}} \\ \mathbf{F}_2 = \mathbf{X}_b\mathbf{B}_{2S} + \mathbf{A}^{\mathrm{T}}\mathbf{G}\left(\mathbf{R}_2^{-1} + \mathbf{R}_2^{-1\mathrm{T}}\right)\mathbf{D}^{\mathrm{T}} + \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}_b\mathbf{D}\mathbf{R}_{2S}\mathbf{D}^{\mathrm{T}} \end{cases}. \quad (13)$$

with $\mathbf{G}=\mathbf{Y}-\mathbf{A}\mathbf{X}_b\mathbf{D}$. The analysis can be written as the solution to Eq. (11) plus the background as follows:

$$\mathbf{X}_a=\mathbf{X}_b+\mathbf{\Delta}. \quad (14)$$

In the ideal case, $\mathbf{A}$ and $\mathbf{D}$ are the identity matrices. So the Eqs. (11-13) can be written as the SE as follows:

$$\left(\mathbf{B}_{\mathbf{R}1} + \mathbf{B}_{\mathbf{R}1}^{\mathrm{T}}\right)\mathbf{\Delta} + \mathbf{\Delta}\left(\mathbf{B}_{\mathbf{R}2} + \mathbf{B}_{\mathbf{R}2}^{\mathrm{T}}\right) - \left(\mathbf{R}_1^{-1\mathrm{T}} + \mathbf{R}_1^{-1}\right)\mathbf{G} - \mathbf{G}\left(\mathbf{R}_2^{-1\mathrm{T}} + \mathbf{R}_2^{-1}\right) = \mathbf{0}, \quad (15)$$

with $\mathbf{B}_{\mathbf{R}1} = \mathbf{B}_1^{-1} + \mathbf{R}_1^{-1}$ and $\mathbf{B}_{\mathbf{R}2} = \mathbf{B}_2^{-1} + \mathbf{R}_2^{-1}$. Compared with the following vector form:

$$(\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})\mathbf{\Delta}-\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{g}=\mathbf{0} \quad (16)$$

with $\mathbf{g}=\mathbf{y}-\mathbf{H}[\mathbf{x}_b]$, the following matrix form is with vector form in $\mathbf{\Delta}$:

$$\left[\mathbf{I}\otimes\left(\mathbf{B}_{\mathrm{R}1} + \mathbf{B}_{\mathrm{R}1}^{\mathrm{T}}\right) + \left(\mathbf{B}_{\mathrm{R}2} + \mathbf{B}_{\mathrm{R}2}^{\mathrm{T}}\right)^{\mathrm{T}} \otimes \mathbf{I}\right]\mathrm{vec}\left(\mathbf{\Delta}\right) - \mathrm{vec}\left[\left(\mathbf{R}_1^{-1\mathrm{T}} + \mathbf{R}_1^{-1}\right)\mathbf{G} + \mathbf{G}\left(\mathbf{R}_2^{-1\mathrm{T}} + \mathbf{R}_2^{-1}\right)\right] = \mathbf{0}. \quad (17)$$

For the case that errors made at different locations are uncorrelated and the variances (means) of the errors are same, Eq. (17) are similar to the Eq. (16) with a constant *k*:

[Eq. (17)]=*k*[Eq. (16)],

as all error covariance matrices in Eq. (17) are the identity matrices multiplied by the estimation of the variance of the errors, which means Eq. (15) is equivalent to Eq. (17).

3. Numerical experiments

3.1 Numerical Experiments in the Ideal Case

Figure 1a shows the ideal field with a dimension of 100×100. Figure 1b shows the background that is generated by the ideal field plus a normal distribution error (the background $\mathbf{x}_b$ is generated by the command xb=peaks(100)+normrnd(μ,σ,m,n) in MATLAB with μ=1, σ=$\sqrt{0.3}$, and the observation is generated in the same way with μ=0 and σ=rand(100)*sqrt(2/10). Here, peaks is a function of two variables, obtained by translating and scaling Gaussian distributions: $z=3(1-x)^2\mathrm{e}^{-x^2-(y+1)^2}-10\left(\frac{x}{5}-x^3-y^5\right)\mathrm{e}^{-x^2-y^2}-\frac{1}{3}\mathrm{e}^{-(x+1)^2-y^2}, -3\le x,y\le 3$, and peaks(100) returns a 100×100 matrix. The function normrnd(μ,σ,m,n) generates random numbers from the normal distribution with mean parameter μ and standard deviation parameter σ, where scalars m and n are the row and column dimensions of output. The function rand(100) returns a 100×100 matrix containing pseudorandom values drawn from the standard uniform distribution on the open interval (0,1). The function sqrt returns the square root of input). The maximum absolute error (MAE) is 2.5260, the root mean squared error (RMSE) is 0.6692, the mean absolute error (MAE1) is 0.5331, and the mean error (ME) is 0.0011. Figure 1c shows the observation. The MAE is 2.9656, the RMSE is 0.7176, the MAE1 is 0.5759 and the ME is –0.0015. Figure 1d shows the analysis that is obtained by Eq. (15). The MAE is 2.0902, the RMSE is 0.4909, the MAE1 is 0.3908, and the ME is −0.0001. The results show that the analysis field provided by Eq. (15) is acceptable in practice. In order to compare Eq. (15) and (16), the experiment (table 1 shows the comparison) with a dimension of 40×40 was established. Both results show that the effect of Eq. (15) is acceptable.

Table 1. The numerical experiment's results (with a dimension of 40×40)

Fields	MAE	RMSE	MAE1	ME
background field	3.0050	0.8969	0.7132	−0.0221
observation field	3.0241	0.8974	0.7215	−0.0117
analysis field (Eq. (15))	2.1370	0.6061	0.4828	−0.0172
analysis field (Eq. (16))	2.4254	0.6977	0.5538	−0.0092

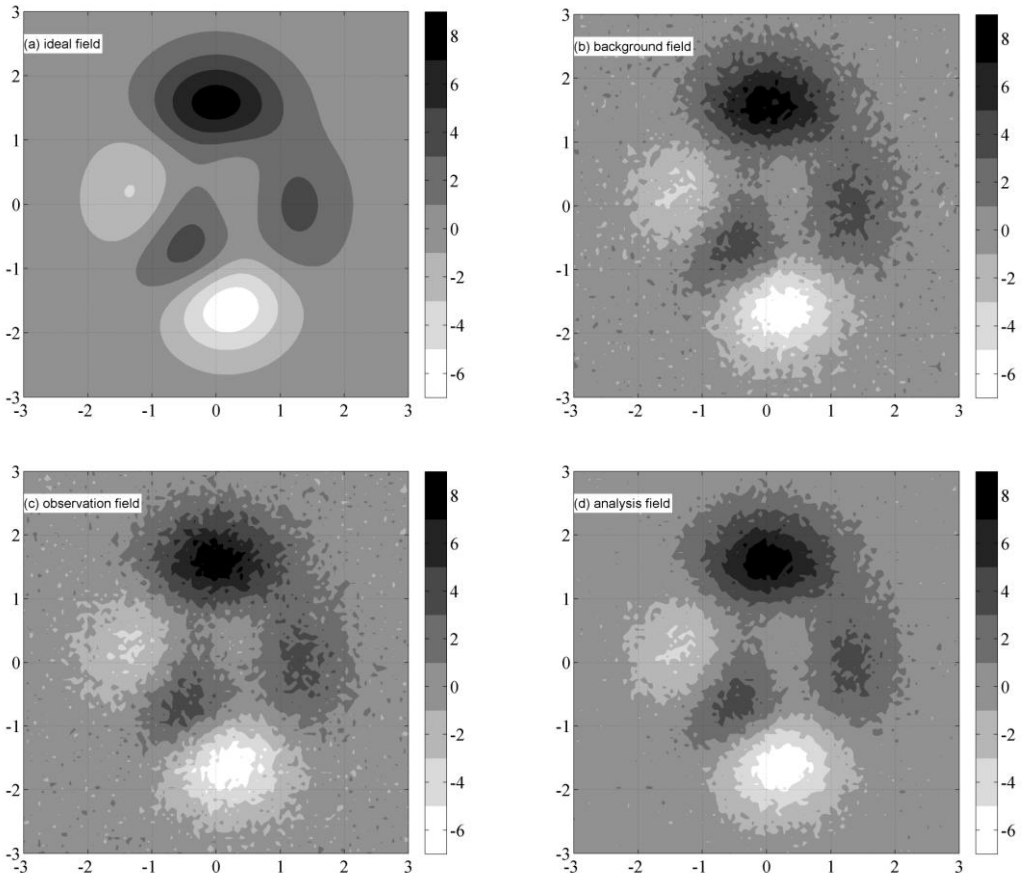

Figure 1. The numerical experiment's results (the ideal case). (a) The ideal field, (b) the background, (c) the observation, and (d) the analysis by Eq. (15).

3.2 Numerical Experiments in the General Case

In the general case, the observational stations cannot be at each the computational grid point. Therefore, there are two ways to resolve this situation. The first way is to interpolate the background onto the observation stations. Contrary to

the first way, the second way is to interpolate the observations onto the computational grid.

The second way seems easier with respect to Eq. (15). The resulting interpolation function is given as follows:

$$p(x_0,y_0)=\begin{cases}\dfrac{\sum_{i=1}^{q}\dfrac{f(x_i,y_i)}{d_i^2}}{\sum_{i=1}^{q}\dfrac{1}{d_i^2}}, d_i=\sqrt{(x_i-x_0)^2+(y_i-y_0)^2}\neq 0\\ f(x_i,y_i), if.d_i=0\end{cases}, \quad (18)$$

where $p(x_0,y_0)$ is the estimation of the position (x_0,y_0), $f(x_i,y_i)$ is the observation of the position (x_i,y_i), d_i is the distances that are arranged from small to large, and q is a constant. Using Eq. (18), we can get the statistical feature values of the observation at each computational grid point.

Figure 2a shows the observation stations that are generated by random numbers. The position of each station is marked with an asterisk. The total number of sites is 900, and the total number of computational grid points is 2500.

In the numerical experiments, the observed error distribution of each station obeys a normal distribution. The mean is 0, and the variance of each point is a random number from 0 to 1. The error of the background field in each grid obeys a normal distribution. The mean value is 1, and the variance is a constant number that belongs to the interval (0,1).

The estimation of the mean of the observed error at each computational point is 0. The mean values estimated distribution from Monte Carlo simulations is shown in

figure 2b, and the comparison of the estimation of the variance is shown in figure 2c. The others are the same as in figure 1 and the comparison is shown in table 2.

Table 2. The numerical experiment's results (with a dimension of 50×50)

Fields	MAE	RMSE	MAE1	ME
background field	2.8215	0.5477	1.0101	0.9960
observation field	1.2196	0.2600	0.1816	−0.0008
analysis field (Eq. (15))	1.1933	0.2078	0.1479	0.0052
analysis field (Eq. (16))	2.1347	0.5144	0.4024	−0.0194

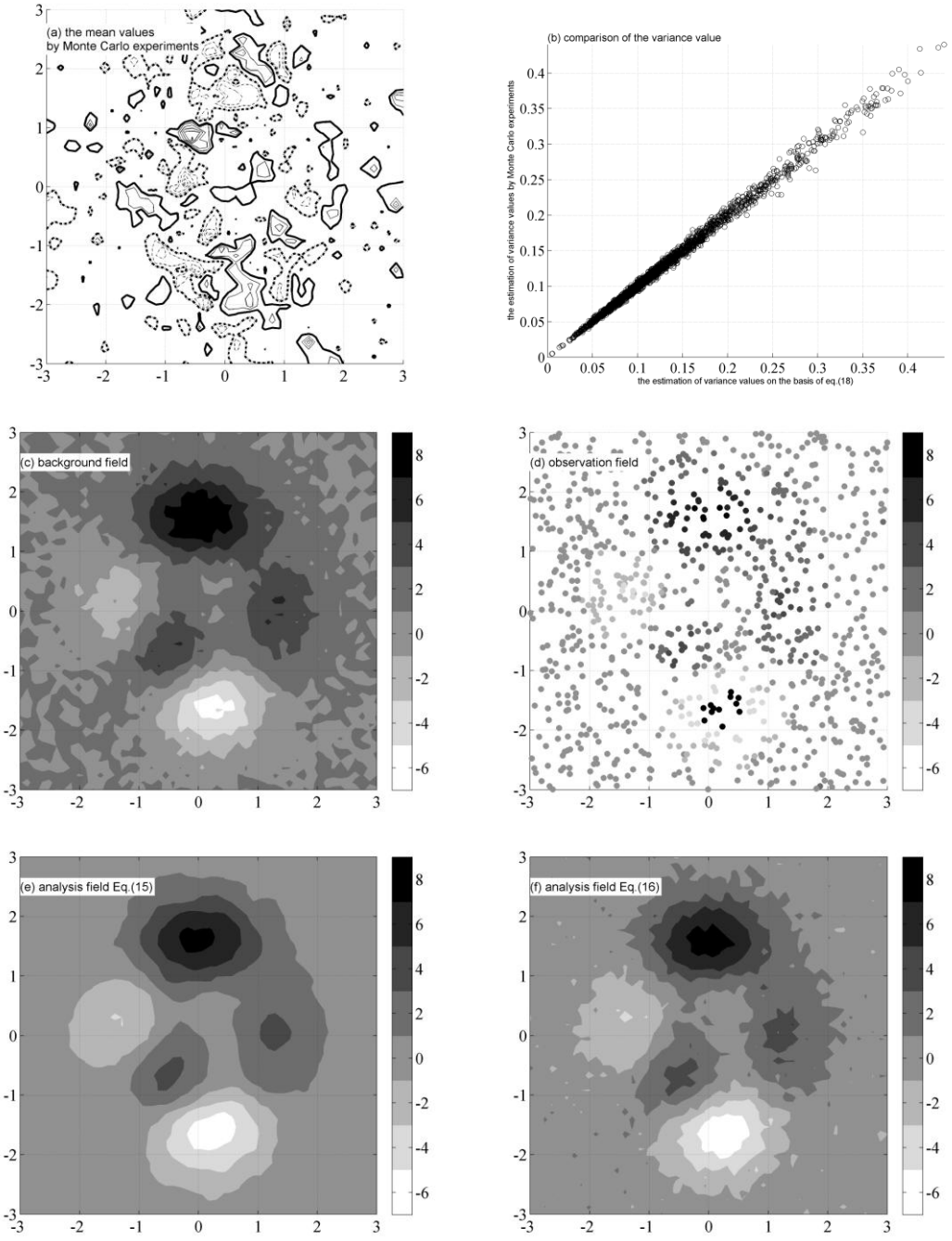

Figure 2. The numerical experiment's results (the general case). (a) the estimation of the mean values (v_m) in each computational point: v_m>0.2 for solid contours, v_m<–0.2 for dashed contours, v_m=0.2 for thick solid contours, v_m=−0.2 for thick dashed contours, and the contour interval is 0.2; (b) the comparison between the estimation of the observed error's variance using Eq. (18) and Monte Carlo experiments; (c) the background; (d) the observation; (e) the analysis of Eq. (15); and (f) the analysis of Eq. (16).

By comparing the analysis field by Eqs. (15) and (16), the MAE, RMSE and MAE1 are smaller for the case with Eq. (15), which means that the errors are concentrated in a smaller interval: for this case in table 2. as an example, the maximum error is 1.1933, but another is 2.1347. The reason that the case in table 2. is better than both case in table 1 and another case in table 2. may be more information about each grid are taken into consideration than others. As the information of each grid are same, the cases in the table 1. are close to each other, but in the case with Eq. (15), the errors are concentrated in a smaller interval, too.

The mean values by Eq.(18) is close to 0 in the figure 2a as the values of the most areas in the figure are on the interval (−0.2 0.2), the variance values of the observation and values gotten by Eq.(18) are mainly on the diagonal line in the scatter plot figure 2b, which means both values are similar. The results that are shown in both figures imply that the estimations of the statistical feature values are acceptable. Both figures 2c-g and the comparison results above show that the effects of this reduction algorithm are no worse than those of the directive scheme for 3DVAR.

4. Summary and discussion

With the numerical results from section 3, the new cost function for 3DVAR within the SE significantly reduces the computational magnitude. The dimension of the covariance matrices $\mathbf{B}_i$ and $\mathbf{R}_i$ (i=1,2) in Eq. (6) is approximately $2\times(m^2+n^2)$, and the dimension of the background error covariance matrix $\mathbf{B}$ in Eq. (3) is approximately $m^2\times n^2$. The rate of both is $2\times(m^2+n^2)/(m^2\times n^2)$, which is a decreasing function of both m and n. For m=n=10, the rate is 0.04, which means that there is an approximately 96% reduction. Without the loss of generality, by setting $m\leqslant n$, the rate takes the following form of an inequality:

$$\frac{4}{n^2}\le\frac{2\left(m^2+n^2\right)}{m^2n^2}=\frac{2\frac{m^2}{n^2}+2}{m^2}\le\frac{4}{m^2},n\ge m>2. \tag{19}$$

Eq. (19) means that there is an approximate reduction of between $4/n^2$ and $4/m^2$. For a high-resolution limited area model or a global model with m, n>100, there is approximately a 99% reduction in and less information for the covariance matrices loss.

Acknowledgments. This work was supported by the Open Research Fund of State Key Laboratory of Estuarine and Coastal Research (Grant number SKLEC-KF201707) and the National Nature Science Foundation of China (Nos. 41490642, 51709243, 51739010, and 41775027).

References

Bannister R. N., 2008a: A review of forecast error covariance statistics in atmospheric

variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134, 1951-1970.

Bannister R. N., 2008b: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, 134, 1971-1996.

Cao Yanhua, Zhu Jiang, Navon I. M., and Luo Zhengdong, 2007: A Reduced-order Approach to Four-dimensional Variational Data Assimilation Using Proper Orthogonal Decomposition. International Journal for Numerical Methods in Fluids, 53, 1571–1583.

Chen Lianglv, Chen Jing, Xue Jishan, and Xia Yu, 2015: Study On the GRAPES Regional En-3DVAR Hybrid Data Assimilation System. *Journal of Meteorological Research*, 29, doi: 10.1007/s13351-015-5021-y.

Chen Xuan, Zheng Chongwei, Zhang Weitao, Jin Peng, and Li Xin, 2017: More Generalized Linear Regression Model and Its Meteorological Application. Journal of PLA University of Science and Technology (Natural Science Edition), 18(2), 144–149. (in Chinese)

Chen Xuan, You Xiaobao, Zheng Chongwei, Sun Wei, and Xie Shenglang, 2019: A New Linear Regression Model and Its Application. Chinese Journal of Atmospheric Sciences, 43(2), 389–400. (in Chinese)Cohn S. E., and Todling R., 1996: Appropriate data assimilation schemes for stable and unstable dynamics.

Journal of the Meteorological Society of Japan, 74(1), 63-75.

Deif A. S., Seif N. P., and Hussein S. A., 1995: Sylvester's Equation: Accuracy and Computational Stability. *Journal of Computational and Applied Mathematics*, 61, 1-11.

Fang F., Zhang T., Pavlidis D., Pain C. C., Buchan A. G., and Navon I. M., 2014: Reduced Order Modelling of an Unstructured Mesh Air Pollution Model and Application in 2D/3D Urban Street Canyons. *Atmospheric Environment*, 96, 96-106.

Farrell B. F., and Ioannu P. J., 2010: State Estimation Using a Reduced Order Kalman Filter. *Journal of the Atmospheric Sciences*, 58(23), 3666-3680.

Federico Stefano, 2013: Preliminary results of a data assimilation system. *Atmospheric and Climate Sciences*, 3, 61-72.

Frolov Sergey, Baptista António M., Leen Todd K., Lu Zhegdong, and van der Merwe Rudolph, 2009: Fast data assimilation using a nonlinear Kalman filter and a model surrogate: an application to the Columbia River estuary. *Dynamics of Atmospheres and Oceans*, 48, 16–45.

Hoteit I., and Köhl A., 2006: Efficiency of Reduced-order, Time-dependent Adjoint Data Assimilation Approaches. *Journal of Oceanography*, 62, 539–550.

Kalnay Eugenia. 2003: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, New York, 168-175.

Kleist Daryl T., Parrish David F., Derber John C., Treadon Russ, Errico Ronald M., and Yang Runhua, 2008: Improving Incremental Balance in the GSI 3DVAR

Analysis System. *Monthly Weather Review*, 137, 1046-1060.

Lawless A. S., Nichols N. K., Boess C., and Bunse-Gerstner A., 2008: Using Model Reduction Methods within Incremental Four-dimensional Variational Data Assimilation. *Monthly Weather Review*, 136, 1511–1522.

Liu Juanjuan, Wang Bin, and Wang Shudong, 2010: The Structure of Background-error Covariance in a Four-dimensional Variational Data Assimilation System: Single-point Experiment. *Advances in Atmospheric Sciences*, 27(6), 1303-1310.

Lorenc A. C., 1986: Analysis Methods for Numerical Weather Prediction. Quarterly *Journal of the Royal Meteorological Society*, 112, 1177–1194.

Mogensen Kristian S., Balmaseda Magdalena Alonso, Weaver Anthony, Martin Matthew, and Vidard Arthur, 2009: NEMOVAR: A Variational Data Assimilation System for the NEMO Ocean Model. *The Meteorology Section of ECMWF Newsletter*, 120, 17-21.

Parrish D. F., and Derber J. D., 1992: The National Meteorological Center spectral Statistical Interpolation Analysis System. *Monthly Weather Review*, 120, 1747-1763.

Robert C., Durbiano S., Blayo E., Verron J., Blum J., and Le Dimet F.X., 2005: A Reduced Order Strategy for 4D-Var Data Assimilation. *Journal of Marine Systems*, 57, 70–82.

Shen Si, Liu Juanjuan, and Wang Bin, 2014: An Extension of the Dimension-reduced Projection 4DVar. *Atmospheric and Oceanic Science Letters*, 7(4), 324-329.

Simoncini V., 2013: Computational Methods for Linear Matrix Equation. http://www.dm.unibo.it/~simoncin/matrixeq.pdf.

Warner Thomas Tomkins, 2011: Numerical Weather and Climate Prediction. Cambridge University Press, New York, 220-233.

Zhao Ying, Wang Bin, and Liu Juanjuan, 2012: A DRP–4DVar Data Assimilation Scheme for Typhoon Initialization Using Sea Level Pressure Data. *Monthly Weather Review*, 140, 1191-1202.