Acta Oceanol. Sin., 2021, Vol. 40, No. 1, P. 135–146
https://doi.org/10.1007/s13131-021-1701-x
http://www.hyxb.org.cn
E-mail: hyxbe@263.net

Study and application of an improved four-dimensional variational assimilation system based on the physical-space statistical analysis for the South China Sea

Yumin Chen[1, 2], Jie Xiang[1], Huadong Du[1*], Sixun Huang[1, 3], Qingtao Song[4]

[1] College of Meteorology and Oceanology, National University of Defense Technology, Nanjing 211101, China

[2] The 93056 Army of People's Liberation Army, Anshan 114000, China

[3] State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou 310012, China

[4] National Satellite Ocean Application Service, Beijing 100081, China

Received 11 February 2020; accepted 11 March 2020

Abstract

The four-dimensional variational assimilation (4D-Var) has been widely used in meteorological and oceanographic data assimilation. This method is usually implemented in the model space, known as primal approach (P4D-Var). Alternatively, physical space analysis system (4D-PSAS) is proposed to reduce the computation cost, in which the 4D-Var problem is solved in physical space (i.e., observation space). In this study, the conjugate gradient (CG) algorithm, implemented in the 4D-PSAS system is evaluated and it is found that the non-monotonic change of the gradient norm of 4D-PSAS cost function causes artificial oscillations of cost function in the iteration process. The reason of non-monotonic variation of gradient norm in 4D-PSAS is then analyzed. In order to overcome the non-monotonic variation of gradient norm, a new algorithm, Minimum Residual (MINRES) algorithm, is implemented in the process of assimilation iteration in this study. Our experimental results show that the improved 4D-PSAS with the MINRES algorithm guarantees the monotonic reduction of gradient norm of cost function, greatly improves the convergence properties of 4D-PSAS as well, and significantly restrains the numerical noises associated with the traditional 4D-PSAS system.

Key words: four-dimensional variational data assimilation (4D-Var), physical space analysis system (PSAS), conjugate gradient algorithm (CG), minimal residual algorithm (MINRES), South China Sea

Citation: Chen Yumin, Xiang Jie, Du Huadong, Huang Sixun, Song Qingtao. 2021. Study and application of an improved four-dimensional variational assimilation system based on the physical-space statistical analysis for the South China Sea. Acta Oceanologica Sinica, 40(1): 135–146, doi: 10.1007/s13131-021-1701-x

1 Introduction

In meteorology and oceanography, variational data assimilation (VDA) uses all valuable observational information to achieve the most accurate description of atmospheric and oceanic states by minimizing the difference between model solutions and observations. The VDA includes three-dimensional VDA (3D-Var) and four-dimensional VDA (4D-Var) (Courtier, 1997; Wang et al., 2019b). Both 3D-Var and 4D-Var play an important role in numerical prediction, reanalysis data construction, parameter inversion and sensitivity analysis (Huang et al., 2005; Moore et al., 2011a, b; Zhang et al., 2017; Zhou et al., 2018; Shi et al., 2018; Wang et al., 2019a).

According to the difference in solution spaces, the 4D-Var can be divided into two kinds. One is solved in the model space, denoted $\mathbb{R}^n$, with the dimension n up to 10^7–10^9, which is usually known as the primal 4D-Var (P4D-Var) (Parrish and Derber, 1992) or the primal formulation of incremental strong constraint 4D-Var (I4D-Var) (Moore et al., 2011c), and the other in the physical space, i.e., the observation space, denoted $\mathbb{R}^m$, with the dimension m only to 10^5, which is usually known as the physical space analysis system (4D-PSAS) (Cohn et al., 1998). Since there is generally $m \ll n$ in atmosphere and ocean, 4D-PSAS is in fact solved in the lower dimensional space, which makes 4D-PSAS have the following two benefits: less memory requirement and lower computational burden; being suitable for application to the weakly constrained 4D-Var with model errors (Trémolet, 2007; Zhong et al., 2012).

Despite the benefits mentioned above, 4D-PSAS is seldom adopted in the atmospheric and oceanic numerical models. While, P4D-Var is now widely used in the 4D-Var system of the European Centre for Medium-Range Weather Forecasts (ECMWF) (Rabier et al., 2000), National Centers for Environmental Prediction (NECP) (Parrish and Derber, 1992), Met Office (Rawlins et al., 2007), China Meteorological Administration (CMA), National Climate Center (NCC) (Liu et al., 2005) and other institutions, and its utility has been widely tested.

Recently, Moore et al. (2011a) develop the Regional Ocean Modeling System (ROMS). The ROMS contains its own 4D-Var system, which includes three modules: a primal formulation of incremental strong constraint 4D-Var (I4D-Var), a dual formula-

Foundation item: The National Key Research and Development Program of China under contract Nos 2017YFC1501803 and 2018YFC1506903; the National Natural Science Foundation of China under contract Nos 91730304, 41475021 and 41575026.
*Corresponding author, E-mail: huadong.du@gmail.com

tion based on a physical-space statistical analysis system (4D-PSAS), and a dual formulation representer-based variant of 4D-Var (R4D-Var). Since 4D-PSAS is seldom applied both in meteorology and in oceanography, the present paper utilizes the ROMS 4D-PSAS for data assimilation (DA) in the South China Sea. This paper is organized as follows: The ROMS and its 4D-Var systems are introduced in Section 2; Section 3 is the application of the ROMS 4D-Var systems in the South China Sea, and both P4D-Var and 4D-PSAS are carried out using the Conjugate gradient algorithm (CG) (Hestenes and Stiefel, 1952); Section 4 is the application of the Minimum residual algorithm (MINRES) (Paige and Saunders, 1975) in 4D-PSAS, and the comparison is made between the CG and MINRES algorithm; and finally, a summary is presented in Section 5.

2 ROMS 4D-Var system

The ROMS is developed by the Institute of Marine and Coastal Sciences of Rutgers University and the University of California, Los Angeles (Shchepetkin and McWilliams, 2005). It is a hydrostatic, primitive equation, Boussinesq ocean general circulation model, widely used in coastal, offshore and marine basins. ROMS uses terrain-following vertical coordinate system (S-coordinate), which allowing for greater vertical resolution in shallow water and regions with complex bathymetry, and can better simulate irregular continental shelf topography. Orthogonal curvilinear coordinates are used allowing for increased horizontal resolution in irregular coastal regions.

The ROMS 4D-Var system includes three modules: I4D-Var, 4D-PSAS and R4D-Var. For each module, ROMS can be used combined with available observations to yield a best estimate of the ocean state. In the primal formulation of I4D-Var the search for the best solution is performed in the full space of the model control vector, while for the dual formulations of 4D-PSAS and R4D-Var only the sub-space of linear functions of the model state vector spanned by the observations (i.e., the dual space) is searched. In oceanographic applications, the number of observations is typically much less than the dimension of the model control vector, and therefore there are remarkable advantages for performing search in the space spanned by the observations. For the completeness of the context and convenience of readers, I4D-Var (i.e., P4D-Var) and 4D-PSAS are summarized as follows.

2.1 *P4D-Var*

In atmospheric and oceanic forecasting, when the model is assumed to be perfect (i.e., the model is error-free), the discrete atmospheric and oceanic numerical models can generally be expressed as follows:

$$\boldsymbol{x}_i = \mathcal{M}_{i,i-1}\boldsymbol{x}_{i-1}, \tag{1}$$

where x_{i-1} and x_i represent the state vectors at the time i–1 and i, respectively, and $\mathcal{M}$ represents the non-linear prediction model.

4D-Var obtains the optimal estimation of the initial field by constructing a cost function $\mathcal{J}$ about the initial field and solving the minimization problem of $\mathcal{J}$ (Kalnay, 2003; Bennett, 2005):

$$\begin{aligned}\mathcal{J}(\boldsymbol{x}_0) =& \frac{1}{2}\left[\boldsymbol{x}_0-\boldsymbol{x}_\mathrm{b}, \boldsymbol{B}^{-1}(\boldsymbol{x}_0-\boldsymbol{x}_b)\right]+\\ &\frac{1}{2}\sum_{i=0}^{M}\left\langle \mathcal{H}_i(\boldsymbol{x}_i)-\boldsymbol{y}_i, \boldsymbol{R}_i^{-1}\left(\mathcal{H}_i(\boldsymbol{x}_i)-\boldsymbol{y}_i\right)\right\rangle = \min!,\end{aligned} \tag{2}$$

where $(\cdot,\cdot)$ represents the inner product in $\mathbb{R}^n$ space, and $\langle\cdot,\cdot\rangle$ represents the inner product in $\mathbb{R}^m$ space. $\boldsymbol{x}_0$ and $\boldsymbol{x}_\mathrm{b}$ are the state vectors of the initial field and background field, the ROMS prognostic variables of a state vector are potential temperature (T), salinity (S), horizontal velocity (u,v), and sea surface displacement (ζ). $\mathcal{H}$ is the non-linear observation operator. $\boldsymbol{B}$ is the error covariance matrix of the background field, in which the spatial correlation operator can be used to adjust the state for one variable in three-dimensional space with the observation of that variable at one location, and the balance operator between different variables can be used to adjust the states for variables with the observations of other variables (Du et al., 2016). M is the time length of observation, $\boldsymbol{y}_i$ and $\boldsymbol{R}_i$ are the observation vector and the observation error covariance matrix at each moment, respectively.

2.1.1 *The formulation in the incremental method*

The dimension of the state vector is very large, up to the order of 10^7–10^9. In order to reduce the computation cost, the incremental method is introduced to VDA by Courtier et al. (1994) through linearizing the nonlinear model and observation operator with respect to the model trajectory (Courtier et al., 1994):

$$\begin{cases}\Delta\boldsymbol{x}_i = \boldsymbol{M}_{i,i-1}\Delta\boldsymbol{x}_{i-1}, & \boldsymbol{M}_{i,i-1} = \left.\dfrac{\partial\mathcal{M}_{i,i-1}}{\partial x}\right|_{\boldsymbol{x}=\boldsymbol{x}_\mathrm{b}},\\ \mathcal{H}_i(\boldsymbol{x}_i) = \mathcal{H}_i(\boldsymbol{x}_\mathrm{b}) + \boldsymbol{H}_i\Delta\boldsymbol{x}_i, & \boldsymbol{H}_i = \left.\dfrac{\partial\mathcal{H}_i}{\partial\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_\mathrm{b}}\end{cases}, \tag{3}$$

where $\Delta\boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{x}_\mathrm{b}$ is state vector increment. Thus, the state vector increment at any time i can be expressed as:

$$\begin{aligned}\Delta\boldsymbol{x}_i &= \boldsymbol{M}_{i,i-1}\Delta\boldsymbol{x}_{i-1}\\ &= \boldsymbol{M}_{i,i-1}\cdot\boldsymbol{M}_{i-1,i-2}\Delta\boldsymbol{x}_{i-2} = \cdots = \boldsymbol{M}_{i,0}\Delta\boldsymbol{x}_0,\end{aligned} \tag{4}$$

where $\boldsymbol{M}_{i,0} = \boldsymbol{M}_{i,i-1}\cdots\boldsymbol{M}_{1,0}$. Therefore, Eq. (2) can be simplified as follows:

$$J(\Delta\boldsymbol{x}_0) = \frac{1}{2}\left(\Delta\boldsymbol{x}_0, \boldsymbol{B}^{-1}\Delta\boldsymbol{x}_0\right) + \frac{1}{2}\left\langle\boldsymbol{G}\Delta\boldsymbol{x}_0-\boldsymbol{d}, \boldsymbol{R}^{-1}\left(\boldsymbol{G}\Delta\boldsymbol{x}_0-\boldsymbol{d}\right)\right\rangle, \tag{5}$$

where $\boldsymbol{d}=[\boldsymbol{d}_1,\cdots,\boldsymbol{d}_M]^\mathrm{T}$ is the innovation vector, $\boldsymbol{d}_i=\boldsymbol{y}_i-\mathcal{H}_i(\boldsymbol{x}_i)$, representing the deviation between the value of model field mapped to the physical space and the observation field at the same time, $\boldsymbol{R}=\mathrm{diag}[\boldsymbol{R}_i,\cdots,\boldsymbol{R}_N]$ is the observation error covariance matrix, $\boldsymbol{G}=[\boldsymbol{G}_0,\cdots,\boldsymbol{G}_M]^\mathrm{T}$ is the generalized observation operator, which is a matrix connecting $\boldsymbol{HM}$ operator, and $\boldsymbol{G}_i=[\boldsymbol{H}_i\boldsymbol{M}_{i,0}]_{m\times n}$.

2.1.2 *The gradient of the cost function and optimal solution*

For P4D-Var, the gradient of J is:

$$\nabla J(\Delta\boldsymbol{x}_0) = \left(\boldsymbol{B}^{-1}+\boldsymbol{G}^\mathrm{T}\boldsymbol{R}^{-1}\boldsymbol{G}\right)\Delta\boldsymbol{x}_0 - \boldsymbol{G}^\mathrm{T}\boldsymbol{R}^{-1}\boldsymbol{d}. \tag{6}$$

According to $\nabla J(\Delta\boldsymbol{x}_0)=0$, the analytic solution for minimization Eq. (5) is as follows:

$$\Delta\boldsymbol{x}_0^a = \left(\boldsymbol{B}^{-1}+\boldsymbol{G}^\mathrm{T}\boldsymbol{R}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^\mathrm{T}\boldsymbol{R}^{-1}\boldsymbol{d}. \tag{7}$$

2.1.3 *Pre-conditioning of Eq. (5)*

In order to avoid the difficulty of inversion of $\boldsymbol{B}$ matrix and improve the condition number of Hessian matrix of cost function, the variable transformation, $\boldsymbol{v}=\boldsymbol{B}^{-1/2}\Delta\boldsymbol{x}_0$, is introduced for

preconditioning (Lorenc, 1988; Tshimanga et al., 2008). Equation (5) can be rewritten as:

$$J(\boldsymbol{v}) = \frac{1}{2}(\boldsymbol{v},\boldsymbol{v}) + \frac{1}{2}\left\langle \boldsymbol{GB}^{1/2}\boldsymbol{v} - \boldsymbol{d}, \boldsymbol{R}^{-1}\left(\boldsymbol{GB}^{1/2}\boldsymbol{v} - \boldsymbol{d}\right)\right\rangle. \quad (8)$$

Then, the gradient and Hessian matrix of J with respect to $\boldsymbol{v}$ are:

$$\nabla_v J = \left(\boldsymbol{I}_n + \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B}^{1/2}\right)\boldsymbol{v} - \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{d}, \quad (9)$$

$$\nabla_v^2 J = \boldsymbol{I}_n + \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B}^{1/2}, \quad (10)$$

where $\boldsymbol{I}_n$ is the unit matrix in the model space. The Hessian matrix refers to the second derivative of the cost function, which determines the properties of the cost function (Thépaut and Moll, 1990).

2.2 *4D-PSAS*

2.2.1 *Transforming the problem solving in model space into physical space*

In order to turn P4D-Var into 4D-PSAS, the analytic solution of the minimization Eq. (7) is, using the Sherman-Morrison-Woodbury Formula $\left(\boldsymbol{B}^{-1}+\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1} = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}})^{-1}$, rewritten as (Amodei, 1995):

$$\Delta\boldsymbol{x}_0^a = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\left(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)^{-1}\boldsymbol{d}. \quad (11)$$

To avoid the inversion of $\left(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)$ matrix, set $\boldsymbol{w} = (\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}})^{-1}\boldsymbol{d}$. Therefore the following linear equations of the $m\times m$ matrix is first solved in the physical space:

$$\left(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)\boldsymbol{w} = \boldsymbol{d}. \quad (12)$$

Then, the solution $\boldsymbol{w}$ is transformed to the state vector increment in the model space, $\Delta\boldsymbol{x}_0 = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}w$.

Equation (12) is equivalent to a minimization problem of the following cost function F:

$$F(\boldsymbol{w}) = \frac{1}{2}\left\langle \boldsymbol{w}, \left(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)\boldsymbol{w}\right\rangle - \langle \boldsymbol{w}, \boldsymbol{d}\rangle, \quad (13)$$

which is usually called the auxiliary cost function in 4D-PSAS (defined in the observation space) so as to distinguish from the cost function J in P4D-Var (defined in the model space).

2.2.2 *Pre-conditioning of Eq. (13)*

By making the variable transformation $\boldsymbol{u} = \boldsymbol{R}^{1/2}\boldsymbol{w}$ for preconditioning (Amodei, 1995), Eq. (13) can be rewritten as:

$$F(\boldsymbol{u}) = \frac{1}{2}\left\langle \boldsymbol{u}, \boldsymbol{R}^{-1/2}\left(\boldsymbol{R}+\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)\boldsymbol{R}^{-1/2}\boldsymbol{u}\right\rangle - \left\langle \boldsymbol{u}, \boldsymbol{R}^{-1/2}\boldsymbol{d}\right\rangle. \quad (14)$$

Then, the gradient and Hessian matrix of F with respect to $\boldsymbol{u}$ are:

$$\nabla_u F = \left(\boldsymbol{I}_m + \boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)\boldsymbol{u} - \boldsymbol{R}^{-1/2}\boldsymbol{d}, \quad (15)$$

$$\nabla_u^2 J = \boldsymbol{I}_m + \boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}, \quad (16)$$

where $\boldsymbol{I}_m$ is the unit matrix in the physical space.

Mapping the solution $\boldsymbol{u}$ of Eq. (14) in 4D-PSAS to the model space, $\tilde{\boldsymbol{v}} = \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{1/2}\boldsymbol{u}$, then the cost function in 4D-PSAS defined in the model space is:

$$\begin{aligned} J_{\mathrm{p}}(\tilde{\boldsymbol{v}}) =& \frac{1}{2}(\tilde{\boldsymbol{v}},\tilde{\boldsymbol{v}}) + \frac{1}{2}\left\langle \boldsymbol{GB}^{1/2}\tilde{\boldsymbol{v}} - \boldsymbol{d}, \boldsymbol{R}^{-1}\left(\boldsymbol{GB}^{1/2}\tilde{\boldsymbol{v}} - \boldsymbol{d}\right)\right\rangle \\ =& \frac{1}{2}\left(\boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{u}, \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{u}\right) + \\ & \frac{1}{2}\left\langle \boldsymbol{GBG}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{u} - \boldsymbol{d}, \boldsymbol{R}^{-1}\left(\boldsymbol{GBG}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{u} - \boldsymbol{d}\right)\right\rangle. \end{aligned} \quad (17)$$

2.3 *Analysis of the equivalence between 4D-PSAS and P4D-Var*

The Hessian matrices of J and F are:

$$\nabla_v^2 J \equiv \boldsymbol{V} = \boldsymbol{I}_n + \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B}^{1/2}, \quad (18)$$

$$\nabla_u^2 F \equiv \boldsymbol{U} = \boldsymbol{I}_m + \boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}. \quad (19)$$

Suppose that the eigenvalue and eigenvector of $\boldsymbol{V}$ is λ_i and $\boldsymbol{V}_i$ respectively, the eigenvalue and eigenvector of $\boldsymbol{U}$ is Λ_i and $\boldsymbol{U}_i$ respectively. Thus,

$$\left(\boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B}^{1/2}\right)\boldsymbol{V}_i = (\lambda_i - 1)\boldsymbol{V}_i, \quad (20)$$

$$\left(\boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)\boldsymbol{U}_i = (\Lambda_i - 1)\boldsymbol{U}_i. \quad (21)$$

Let $\boldsymbol{V}_i = \boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{U}_i$, then Eq. (20) can be rewritten as:

$$\begin{aligned} &\left(\boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)^{-1}\left(\boldsymbol{B}^{1/2}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)\left(\boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)\boldsymbol{U}_i \\ &= (\lambda_i - 1)\boldsymbol{U}_i. \end{aligned} \quad (22)$$

In the special case m=n, $\boldsymbol{V}$ and $\boldsymbol{U}$ have the same eigenvalue, i.e., $\lambda_i = \Lambda_i$. So, the Hessian matrices of J and F have the same condition number $C(\boldsymbol{V}) = C(\boldsymbol{U}) = \lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ and $\lambda_{\min}$ represent the maximum and minimum eigenvalues respectively. In the case $m \ll n$, the singular value decomposition of matrices can be used to prove that the Hessian matrices of J and F have the same condition number.

3 Application of the ROMS 4D-Var system in the South China Sea

The ROMS 4D-Var system is now applied to the case of South China Sea, and the observational data are temperature profile and sea surface temperature. Both P4D-Var and 4D-PSAS are fulfilled and the analysis field by P4D-Var is regarded as a reference. The assimilation utility of 4D-PSAS is analyzed according to the cost function curves and the increment fields obtained by the assimilation. The optimal algorithm used in P4D-Var and 4D-PSAS is the CG algorithm (Moore et al., 2011c).

3.1 *Experiment configuration and data*

3.1.1 *Model domain configuration*

The model domain covers the South China Sea and its adjacent waters (5.44°S–25.36°N, 98.05°–126.73°E). Figure 1 shows the topographic depth in the model domain, provided by the ETOPO1 data (Amante and Eakins, 2009). The maximum depth is set to 5 km, with a horizontal resolution of (1/4)°×(1/4)° and 32 *S*-

coordinate layers in the vertical. The stretching parameters of S-coordinate are set to $\theta_s = 5$ and $\theta_b = 0.4$, that is, increase vertical resolution near the sea surface, and Generic Length-Scale mixing (GLS) scheme is adopted as turbulence closure scheme. On the four boundaries of the model domain, different open boundary conditions are chosen for different physical variables. For example, the model is configured to conserve volume with a free-surface Chapman condition, a Flather condition for 2D momentum, and Clamped condition for 3D momentum and tracers. These lateral boundary conditions are all provided by the simple ocean data assimilation (SODA) data set (Dee et al., 2011). And a sponge layer is also used to each open boundary where the viscosity increased linearly from 4 m^2/s^2 in the interior to 100 m^2/s^2 at the boundary over a distance of about 100 km.

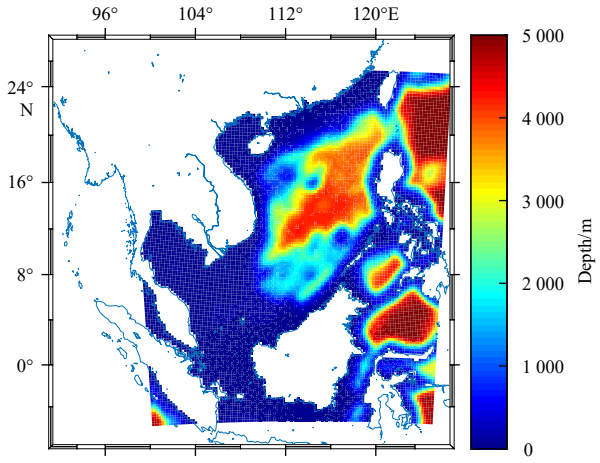

Fig. 1. Topographic depth of model domain.

3.1.2 *Assimilation configuration*

This paper chooses the ocean reanalysis data of Hybrid Coordinate Ocean Model (HYCOM) in the model domain at 00:00 on February 1, 2012 as the initial field (Shu et al., 2014) and NCEP reanalysis data as the forcing field (Amenu and Kumar, 2005). Firstly, the spin up process is carried out, and the model runs for a model year to obtain a stable state of the ocean. The model output field at 00:00 on February 1, 2013 was used as the initial field for DA experiment. The assimilation window was 4 d and the number of iterations was set at 50 times. The observations used is a blend of data from the version 4 of the Met Office Hadley Centre EN series of data sets (EN4) (Good et al., 2013) and the sea surface temperature data of the National Oceanic and Atmospheric Administration (NOAA) (Lauritson, 1991) in the assimilation window. The EN4 data set is a set of global quality-controlled ocean temperature and salinity profiles collected, processed and released by the Met Office Hadley Centre. Sea surface temperature data, the daily and grid product, is provided by the National Climatic Data Center (NCDC) of NOAA.

3.2 *Experiment results and analysis*

3.2.1 *Analysis of the cost function cures*

Figure 2 shows the dependence of $\log_{10}J$ and $\log_{10}J_p$ on number of iterations. It is evident that when the number of iterations reaches 50, $\log_{10}J$ and $\log_{10}J_p$ nearly coincide, and both J and J_p achieve effective convergence to a minimum value. During the iteration process, J decreases monotonically, and after about 35 iterations, J almost reach a constant value, which indicates that J has reached a good convergence. However, the reduction process of J_p has obvious oscillation phenomenon, and converge to the minimum more slowly than J. It is only after 22 iterations that the value of J_p is completely smaller than the initial value. This shows that 4D-PSAS and P4D-Var have the same value of cost function only after the cost functions reach complete convergence.

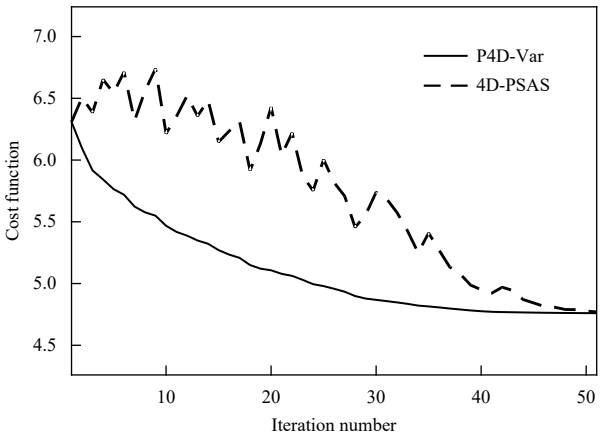

Fig. 2. $\log_{10}J$ and $\log_{10}J_p$ vs number of iterations for data assimilation windows of 4 d duration for a representative assimilation cycle starting on February 1, 2013 using P4D-Var (real line) and 4D-PSAS (dashed line) respectively.

3.2.2 *Analysis of the increment fields*

Figures 3a and b show the sea surface temperature increments (i.e., the deviation between the analysis field and the background field) obtained after the 50th iteration of P4D-Var and 4D-PSAS, respectively. It can be found that the increments obtained by 4D-PSAS and P4D-Var are largely identical when cost functions reach complete convergence after 50 iterations, and the average relative deviation between them is only 0.28% and the root mean square error (RMSE) is 0.50 K.

Figures 3c and d show the sea surface temperature increments obtained after the 10th iteration of P4D-Var and 4D-PSAS, respectively. It can be seen that there are obvious differences, with an average relative deviation of 24.63% and a RMSE is 0.85 K. This shows that the solutions obtained using 4D-PSAS and P4D-Var are, in general, very different if the iterations are terminated before the asymptote in cost functions have been reached.

3.2.3 *Theoretical analysis*

The difference in cost functions between 4D-PSAS and P4D-Var is analyzed firstly from the physical sense. The cost function J of P4D-Var derives from the posterior probability density function, which represents the analysis error. Decrease in J means increase in accuracy of analysis solutions, so iterations in P4D-Var ensures that the analysis solutions are more and more accurate. However, for 4D-PSAS, the auxiliary cost function F has no direct physical meaning, and the iteration solution is often undesirable until the cost function achieves its minimum.

On the other hand, at the point of minimum of F, $\nabla_u F = 0$, yields the analysis increment in physical space as follows:

$$\boldsymbol{u}^{\mathrm{a}} = \left(\boldsymbol{I}_m + \boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\right)^{-1}\boldsymbol{R}^{-1/2}\boldsymbol{d}. \tag{23}$$

At $\boldsymbol{u}^{\mathrm{a}}$, F arrives at its minimum:

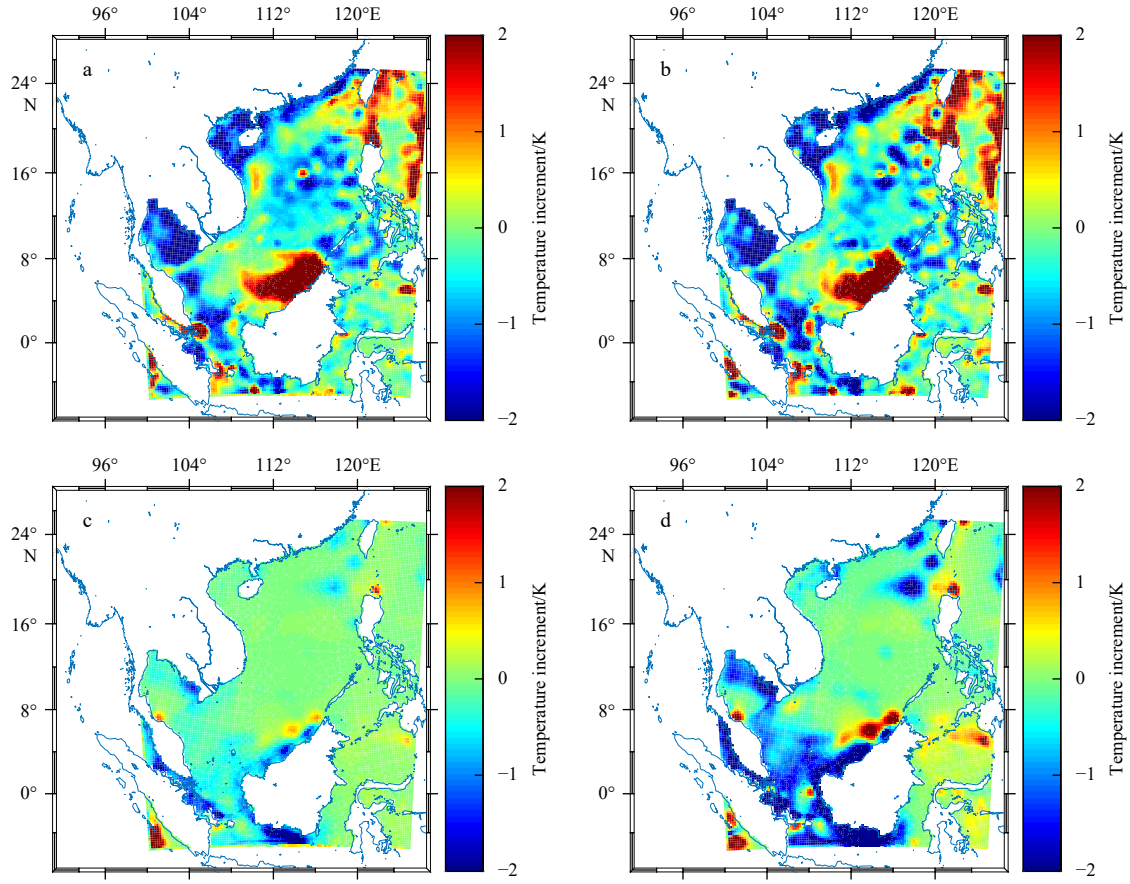

Fig. 3. Sea surface temperature increments obtained using P4D-Var and 4D-PSAS, where calculation is terminated at the 50th and 10th iteration respectively. a. P4D-Var, 50 iterations; b. 4D-PSAS, 50 iterations; c. P4D-Var, 10 iterations; d. 4D-PSAS, 10 iterations.

$$F(\boldsymbol{u}^{\mathrm{a}}) = -\frac{1}{2}\left\langle \boldsymbol{u}^{\mathrm{a}}, \boldsymbol{R}^{-1/2}\boldsymbol{d}\right\rangle. \tag{24}$$

Through the variable transformations, $\Delta\boldsymbol{x}_0 = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{w}$ and $\boldsymbol{u} = \boldsymbol{R}^{1/2}\boldsymbol{w}$, the minimum of J is:

$$J(\Delta\boldsymbol{x}_0) = \frac{1}{2}\left\langle \boldsymbol{u}^{\mathrm{a}}, \boldsymbol{R}^{-1/2}\boldsymbol{d}\right\rangle \equiv -F(\boldsymbol{u}^{\mathrm{a}}). \tag{25}$$

From Eq. (25), it is obvious that when the cost functions reach their minimum, $J(\boldsymbol{v})$ and $F(\boldsymbol{u})$ are of opposite sign. Since the analysis solution in P4D-Var is determined directly by minimizing the cost function $J(\boldsymbol{v})$, while in 4D-PSAS is determined indirectly by minimizing the auxiliary cost function $F(\boldsymbol{u})$. Therefore, the equivalence between 4D-PSAS and P4D-Var can only guarantee the similarity in convergence of $F(\boldsymbol{u})$ and of $J(\boldsymbol{v})$, but cannot guarantee that between $J_{\mathrm{p}}(\tilde{\boldsymbol{v}})$, which is obtained by a transformation from the physical space to the model space, and $J(\boldsymbol{v})$.

Next, we will analyze the relationship between the cost function J_{p} and auxiliary cost function F. According to Eq. (17), using $\boldsymbol{L} = \boldsymbol{R}^{-1/2}\boldsymbol{G}\boldsymbol{B}^{1/2}$, J_{p} can be written as:

$$\begin{aligned} J_{\mathrm{p}}(\tilde{\boldsymbol{v}}) =& \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{u} - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{R}^{-1/2}\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{d} \\ =& \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\left(\boldsymbol{I}_n + \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{L}^{\mathrm{T}}\boldsymbol{u} - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{d}' + \frac{1}{2}\boldsymbol{d}'^{\mathrm{T}}\boldsymbol{d}' \\ =& \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{I}_m + \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\right)\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{u} - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{d}' + \frac{1}{2}\boldsymbol{d}'^{\mathrm{T}}\boldsymbol{d}', \end{aligned} \tag{26}$$

where $\boldsymbol{d}' = \boldsymbol{R}^{-1/2}\boldsymbol{d}$.

Similarly, according to Eq. (14), using the $\boldsymbol{L}$ operator, F can be written as:

$$F(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{I}_m + \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\right)\boldsymbol{u} - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{d}'. \tag{27}$$

Rearranging terms, Eq. (26) can be rewritten as:

$$\begin{aligned} J_{\mathrm{p}}(\tilde{\boldsymbol{v}}) =& \frac{1}{2}\left[\left(\boldsymbol{I}_m + \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\right)\boldsymbol{u} - \boldsymbol{d}'\right]^{\mathrm{T}}\left[\left(\boldsymbol{I}_m + \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\right)\boldsymbol{u} - \boldsymbol{d}'\right] - \\ & \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{I}_m + \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\right)\boldsymbol{u} + \boldsymbol{d}'^{\mathrm{T}}\boldsymbol{u}. \end{aligned} \tag{28}$$

Using Eqs (27) and (28), the relationship between J_{p} and F is:

$$J_{\mathrm{p}}(\tilde{\boldsymbol{v}}) = \frac{1}{2}\left\|\nabla_u F\right\|^2 - F(\boldsymbol{u}). \tag{29}$$

According to Eq. (29), when F is mapped from the physical space to J_{p} in the model space, J_{p} is related not only to F itself, but also to the norm of gradient of F.

Figure 4 shows the dependence of the gradient norm $\|\nabla_u F\|$ on iteration numbers in 4D-PSAS. At the beginning of iteration, F is near zero due to $\boldsymbol{x}_0 = \boldsymbol{x}_{\mathrm{b}}$. So $\|\nabla_u F\|$ overweighs F, and $\|\nabla_u F\|$ oscillates obviously before F achieves its minimum. This is the reason for oscillation of J_{p}, which seriously affects the convergence properties of 4D-PSAS.

By analyzing the principle of optimization algorithm, this paper explores the causes of oscillation of the gradient norm. In the

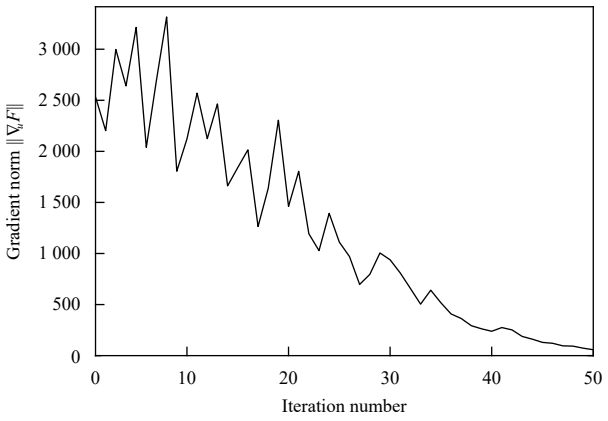

Fig. 4. $\|\nabla_u F\|$ vs number of iterations for data assimilation windows of 4 d duration for a representative assimilation cycle starting on February 1, 2013 using 4D-PSAS.

following section, the feasibility of improving the convergence property of 4D-PSAS using the MINRES algorithm, which guarantees the monotonic reduction of residual error, is studied, and its effect is verified by numerical experiments.

4 Implementation of MINRES algorithm in 4D-PSAS

4.1 *Comparisons of the MINRES and CG algorithms*

Here, for brevity of contents, more about the CG and MINRES algorithms are attached as an appendix. The CG algorithm is applied to solving Eq. (A11) and an approximate solution is found by minimizing the $\boldsymbol{A}$ norm of the absolute error.

$$\boldsymbol{x}_l = \|\boldsymbol{x} - \boldsymbol{x}_*\|_A = \min_{\boldsymbol{x}\in\boldsymbol{x}_0+\mathcal{K}^l}! \,. \tag{30}$$

As stated in Appendix, $\boldsymbol{x}_l$ is the solution of the l-th iteration of Eq. (A11) and $\boldsymbol{x}_*$ is the true solution of Eq. (A11). Equation (30) just makes J or F monotonically reduce in their respective solution space. Therefore, the CG algorithm does not have any constraints to make the gradient norm monotonically reduce, so it cannot guarantee that J_p decreases monotonically.

Unlike the CG algorithm, the MINRES algorithm is applied to solving Eq. (A14), and an approximate solution is found by minimizing the residual norm.

$$\boldsymbol{x}_l = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 = \min_{\boldsymbol{x}\in\boldsymbol{x}_0+\mathcal{K}^l}! \,. \tag{31}$$

According to recursive relation Eq. (A16), $\rho_{l+1} = -\rho_l s_{l+1}$, in which ρ_{l+1} is a new residual error, s_{l+1} is the coefficient of QR decomposition using a Givens transform, satisfying $|s_{l+1}| < 1$. So, the MINRES algorithm can guarantee monotonic reduction of residual norm. And in 4D-Var, the residual norm equals to the gradient norm, so the gradient norm can theoretically decrease monotonically and the convergence property of 4D-PSAS can be improved.

4.2 *Experiments using the MINRES algorithm and analysis of the results in 4D-PSAS*

The improvement of 4D-PSAS using MINRES algorithm is tested from three aspects: gradient norm curves, cost function curves and increment fields. The experiment configuration is consistent with Section 2.4.1.

4.2.1 *Analysis of the gradient norm curves*

Figure 5 shows the dependence of the gradient norm on number of iterations in 4D-PSAS. As can be seen from Fig. 5, the gradient norm obtained by the MINRES algorithm decreases significantly compared with the CG algorithm, and the oscillation phenomenon disappears during the iteration process.

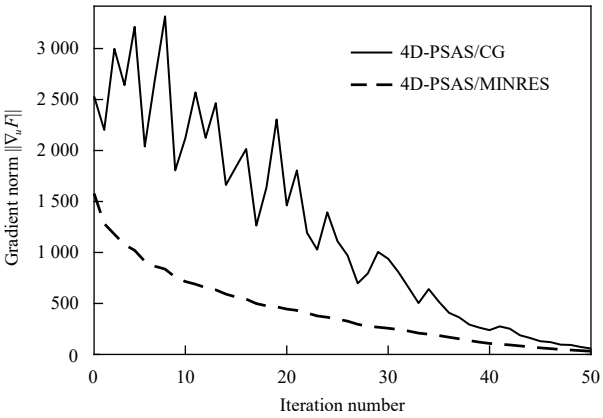

Fig. 5. $\|\nabla_u F\|$ vs number of iterations for data assimilation windows of 4 d duration for a representative assimilation cycle starting on February 1, 2013 using 4D-PSAS/CG (real line) and 4D-PSAS/MINRES (dashed line) respectively.

4.2.2 *Analysis of the cost function curves*

Figure 6 shows the dependence of $\log_{10} J$ on number of iterations in P4D-Var, the dependence of $\log_{10} J_\mathrm{p}$ on number of iterations both in 4D-PSAS/CG and in 4D-PSAS/MINRES. It is clear that, during the iteration process, the cost function curve in 4D-PSAS/MINRES is closer to that in P4D-Var/CG, and the convergence rate of J_p in 4D-PSAS/MINRES is faster than that in 4D-PSAS/CG. Therefore, for 4D-PSAS, the convergence rate of J_p is obviously faster using the MINRES algorithm, and J_p can decrease monotonically, which ensures that the analysis solution in each iteration is better than the previous one before final convergence.

From Fig. 6, it seems that the convergence in P4D-Var/CG is better than that in 4D-PSAS/MINRES. But in fact, the two schemes are for different spaces, the model space and physical space. Usually, the model space and physical space differ by the order 10^2 in dimension, so the computation time for the two schemes should also be considered for comparison. As shown in Fig. 7, the relationship is given between the cost function and computer time over the time interval from 10 000 s to 18 000 s, and it should be noted that the number of iterations for the three experiments were all set to 50, so the three curve were stooped at different time. It can be seen that the final time at complete convergence in 4D-PSAS is about 3 000–4 000 s earlier than that in P4D-Var, which reduce the running time by roughly 12 % in assimilation.

4.2.3 *Analysis of the increment fields*

Figure 8 shows the sea surface temperature increments obtained after 10 iterations of assimilation experiments. Figure 8a is the increment obtained by P4D-Var, and the optimization algorithm used in the iteration is CG algorithm; Figs 8b and c are the increments obtained by 4D-PSAS, and the optimization algorithms used in the iteration are CG algorithm and MINRES algorithm, respectively.

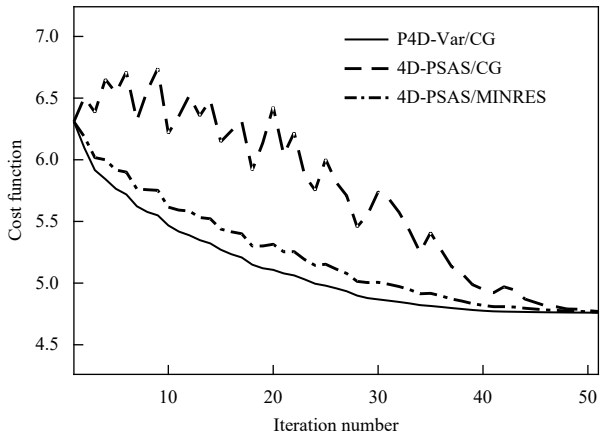

Fig. 6. The dependences of $\log_{10}J$ in P4D-Var/CG (real line), $\log_{10}J_p$ in 4D-PSAS/CG (dashed line), and $\log_{10}J_p$ in 4D-PSAS/MINRES (dash-dotted line) on number of iterations for data assimilation windows of 4 d duration for a representative assimilation cycle starting on February 1, 2013.

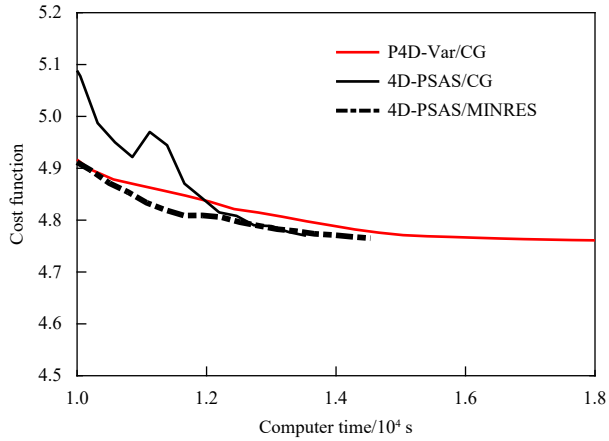

Fig. 7. The dependences of $\log_{10}J$ in P4D-Var/CG (red line), and $\log_{10}J_p$ in 4D-PSAS/CG (black line) and in 4D-PSAS/MINRES (dashed line) on computer times for data assimilation windows of 4 d duration for a representative assimilation cycle starting on February 1, 2013.

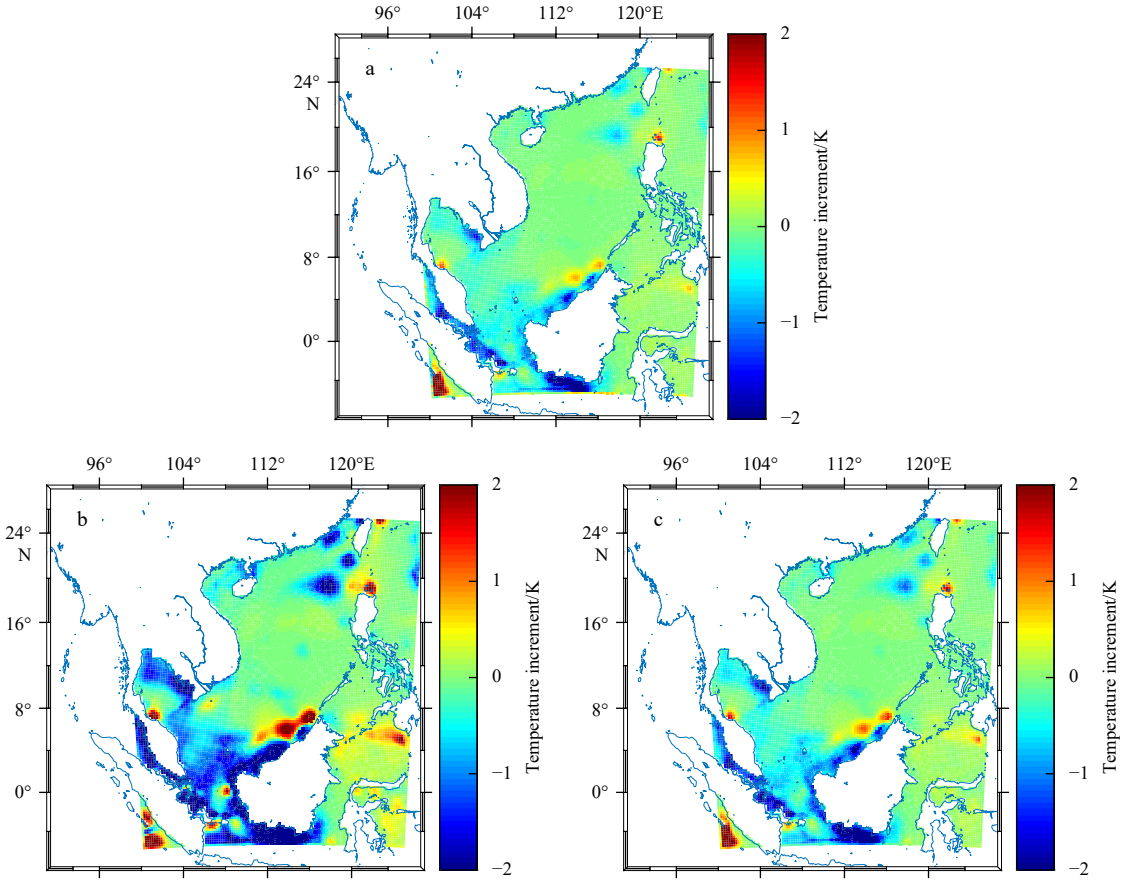

Fig. 8. Sea surface temperature increments after the assimilation, in which the calculation is terminated at 10 iterations. a. P4D-Var/CG; b. 4D-PSAS/CG; c. 4D-PSAS/MINRES.

After using MINRES algorithm, we can see that the increment obtained by 4D-PSAS/MINRES are improved significantly. Compared with results of Section 2.4.2, The average relative deviation between the increment fields of 4D-PSAS and P4D-Var is reduced from 24.63% to 5.04%, and the RMSE is reduced from 0.85 K to 0.19 K. This shows that MINRES algorithm can effectively improve the assimilation results of 4D-PSAS during the iteration process.

5 Summary

In this paper, the ROMS 4D-Var system is applied to data assimilation of South China Sea, and two kinds of schemes, P4D-Var (in the model space) and 4D-PSAS (in the observation space) modules are performed. Theoretically, the 4D-PSAS data assimil-

ation scheme is superior to the P4D-Var scheme. However, when the CG algorithm is employed in P4D-Var and 4D-PSAS respectively, the iteration processes exhibit different characteristics. Although the effective convergence of the cost functions is reached both in P4D-Var and in 4D-PSAS, the behaviors of the cost functions, J and J_p, are different. Obviously, J decreases monotonically, while J_p shows fluctuational property, and the convergence rate of J_p is obviously slower than that of J. In addition, the average relative deviation between the sea surface temperature increments obtained using P4D-Var and 4D-PSAS is 24.63% after the 10th iteration. The non-monotonic reduction of J_p is found to be due to the non-monotonic variation of the gradient norm of the auxiliary cost function F in 4D-PSAS. In order to suppress the non-monotonic variation of the gradient norm of J_p during the iteration process, the CG algorithm is replaced by the MINRES algorithm in 4D-PSAS. The data assimilation experiments indicate that the cost function J_p of 4D-PSAS decreases monotonically and the convergence rate clearly increases. And, the sea surface temperature increment obtained by 4D-PSAS after 10th iteration is basically consistent with that of P4D-Var, and the average relative deviation between the sea surface temperature increments obtained by 4D-PSAS and P4D-Var is reduced from 24.63% to 5.04%.

To sum up, the MINRES algorithm can not only guarantee monotonic reduction of the gradient norm of the cost function J_p in 4D-PSAS, but also can significantly improve the convergence property of the iteration process.

References

Amante C, Eakins B W. 2009. ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24. Boulder, CO: National Geophysical Data Center, Marine Geology and Geophysics Division

Amenu G G, Kumar P. 2005. NVAP and Reanalysis-2 global precipitable water products: Intercomparison and variability studies. Bulletin of the American Meteorological Society, 86(2): 245–256, doi: 10.1175/BAMS-86-2-245

Amodei L. 1995. Approached solution for a data assimilation problem taking into account model errors. Comptes Rendus De L Academie Des Sciences, 321: 1087–1094

Bennett A F. 2005. Inverse Modeling of the Ocean and Atmosphere. Cambridge: Cambridge University Press, 86–93

Cohn S E, Da Silva A, Guo J, et al. 1998. Assessing the effects of data selection with DAO Physical Space Statistical Analysis System. Monthly Weather Review, 126: 2913–2926, doi: 10.1175/1520-0493(1998)126<2913:ATEODS>2.0.CO;2

Courtier P. 1997. Dual formulation of four-dimensional variational assimilation. Quarterly Journal of the Royal Meteorological Society, 123(544): 2449–2461, doi: 10.1002/qj.49712354414

Courtier P, Thépaut J N, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. Quarterly Journal of the Royal Meteorological Society, 120(519): 1367–1387, doi: 10.1002/qj.49712051912

Dee D P, Uppala S M, Simmons A J, et al. 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137(656): 553–597, doi: 10.1002/qj.828

Desroziers G, Berre L. 2012. Accelerating and parallelizing minimizations in ensemble and deterministic variational assimilations. Quarterly Journal of the Royal Meteorological Society, 138(667): 1599–1610, doi: 10.1002/qj.1886

Du Huadong, Zhang Gui, Yang Pinglyu, et al. 2016. Construction of background error covariance matrix in oceanic variational data assimilation. Journal of PLA University of Science and Technology (Natural Science Edition) (in Chinese), 17(1): 72–80

Good S A, Martin M J, Rayner N A. 2013. EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. Journal of Geophysical Research: Oceans, 118(12): 6704–6716, doi: 10.1002/2013JC009067

Hestenes M R, Stiefel E. 1952. Methods of conjugate gradients for solving linear systems. Journal of Research of the National Bureau of Standards, 49(6): 409–436, doi: 10.6028/jres.049.044

Huang Sixun, Xiang Jie, Du Huadong, et al. 2005. Inverse problems in atmospheric science and their application. Journal of Physics: Conference Series, 12: 005

Kalnay E. 2003. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge: Cambridge University Press, 168–175

Lanczos C. 1950. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. Journal of research of the National Bureau of Standards, 45(4): 255–282, doi: 10.6028/jres.045.026

Lauritson L. 1991. The accessibility of satellite data at the National climatic Data Center (NCDC). Global and Planetary Change, 4(1–3): 279–280, doi: 10.1016/0921-8181(91)90106-7

Liu Yimin, Li Weijing, Zhang Peiqun. 2005. A global 4-dimensional ocean data assimilation system and the studies on its results in the tropic Pacific. Acta Oceanologica Sinica (in Chinese), 27(1): 27–35

Lorenc A C. 1988. Optimal nonlinear objective analysis. Quarterly Journal of the Royal Meteorological Society, 114(479): 205–240, doi: 10.1002/qj.49711447911

Moore A M, Arango H G, Broquet G, et al. 2011a. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part Ⅱ. Performance and application to the California Current System. Progress in Oceanography, 91(1): 50–73, doi: 10.1016/j.pocean.2011.05.003

Moore A M, Arango H G, Broquet G, et al. 2011b. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part Ⅲ. Observation impact and observation sensitivity in the California Current System. Progress in Oceanography, 91(1): 74–94, doi: 10.1016/j.pocean.2011.05.005

Moore A M, Arango H G, Broquet G, et al. 2011c. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I. System overview and formulation. Progress in Oceanography, 91(1): 34–49, doi: 10.1016/j.pocean.2011.05.004

Paige C C, Saunders M A. 1975. Solution of sparse indefinite systems of linear equations. SIAM Journal on Numerical Analysis, 12(4): 617–629, doi: 10.1137/0712047

Parrish D F, Derber J C. 1992. The National Meteorological Center's spectral statistical-interpolation analysis system. Monthly Weather Review, 120(8): 1747–1763, doi: 10.1175/1520-0493(1992)120<1747:TNMCSS>2.0.CO;2

Rabier F, Järvinen H, Klinker E, et al. 2000. The ECMWF operational implementation of four-dimensional variational assimilation: I. Experimental results with simplified physics. Quarterly Journal of the Royal Meteorological Society, 126(564): 1143–1170, doi: 10.1002/qj.49712656415

Rawlins F, Ballard S P, Bovis K J, et al. 2007. The Met Office global four-dimensional variational data assimilation scheme. Quarterly Journal of the Royal Meteorological Society, 133(623): 347–362, doi: 10.1002/qj.32

Saad Y. 2003. Iterative Methods for Sparse Linear Systems. 2nd ed. Philadelphia: SIAM, 174–177

Shchepetkin A F, McWilliams J C. 2005. The regional oceanic modeling system (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. Ocean modelling, 9(4): 347–404, doi: 10.1016/j.ocemod.2004.08.002

Shi Junqiang, Yin Xunqiang, Qiao Fangli. 2018. Optimizing the spatial ocean observation system based on data assimilation assessment: The Gulf of Thailand as an example. Haiyang Xuebao (in Chinese), 40(2): 14–29

Shu Yeqiang, Xue Huijie, Wang Dongxiao, et al. 2014. Meridional overturning circulation in the South China Sea envisioned from the high-resolution global reanalysis data GLBa0.08. Journal of Geophysical Research: Oceans, 119(5): 3012–3028, doi: 10.1002/2013JC009583

Thépaut J N, Moll P. 1990. Variational inversion of simulated TOVS

radiances using the adjoint technique. Quarterly Journal of the Royal Meteorological Society, 116(496): 1425–1448, doi: 10.1002/qj.49711649609

Trémolet Y. 2007. Model-error estimation in 4D-Var. Quarterly Journal of the Royal Meteorological Society, 133(626): 1267–1280, doi: 10.1002/qj.94

Tshimanga J, Gratton S, Weaver A T, et al. 2008. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. Quarterly Journal of the Royal Meteorological Society, 134(632): 751–769, doi: 10.1002/qj.228

Wang Yuepeng, Hu Kun, Ren Lanlan, et al. 2019a. Optimal observations-based retrieval of topography in 2D shallow water equations using PC-EnKF. Journal of Computational Physics, 382: 43–60, doi: 10.1016/j.jcp.2019.01.004

Wang Yuepeng, Ren Lanlan, Zhang Zongyuan, et al. 2019b. Sparsity-promoting elastic net method with rotations for high-dimensional nonlinear inverse problem. Computer Methods in Applied Mechanics and Engineering, 345: 263–282, doi: 10.1016/j.cma.2018.10.040

Zhang Kaifeng, Deng Wanyue, Wang Ting, et al. 2017. Blending satellite scatterometer data based on variational with multi-parameter regularization method. Haiyang Xuebao (in Chinese), 39(12): 122–135

Zhong Jian, Fei Jianfang, Huang Sixun, et al. 2012. Study of weak constraint 4dvar with model error forcing control variable. Acta Physica Sinica (in Chinese), 61(14): 149203

Zhou Chaojie, Zhang Xiaohua, Zhang Jie, et al. 2018. An evaluation of sea surface height assimilation using along-track and gridded products based on the Regional Ocean Modeling System (ROMS) and the four-dimensional variational data assimilation. Acta Oceanologica Sinica, 37(9): 50–58, doi: 10.1007/s13131-018-1225-1

Appendix: Introduction of the CG algorithm and MINRES algorithm in the Krylov subspace

A1 Krylov subspace method

Krylov subspace method is a practical and effective method to solve large matrix calculation (Saad, 2003). Especially in recent years, the computational efficiency of this method is getting higher, and its application scope is becoming wider. This method is developed on the basis of Galerkin algorithm. Its purpose is to approximate the solution of high-dimensional problems with the solution of low-dimensional problems, that is, to select an appropriate l-dimensional subspace, $\mathcal{K}^l \in \mathbb{R}^m (l \ll m)$, and then "project" the problem considered to $\mathcal{K}^l$ in a specific sense to become a low-dimensional problem.

For 4D-PSAS, the following equation need to be solved in physical space $\mathbb{R}^m$.

$$\left(\boldsymbol{R} + \boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\right)\boldsymbol{w} = \boldsymbol{d}. \tag{A1}$$

Equation (A1) can be abbreviated as $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}_{m\times m}$ is a symmetric matrix. It is also equivalent to the following minimization problems:

$$J(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle = \min!. \tag{A2}$$

In the iterative solution process of Eq. (A2), if the initial value is $\boldsymbol{x}_0$, the initial residual error is $r_0 = -\nabla_{\boldsymbol{x}} J|_{\boldsymbol{x}_0} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$.

At this point, Krylov space is $\mathcal{K}^l(\boldsymbol{A}, r_0) = \mathrm{span}\left\{r_0, \boldsymbol{A}r_0, \boldsymbol{A}^2 r_0, \cdots, \boldsymbol{A}^{l-1} r_0\right\}$, in which r_0, $\boldsymbol{A}r_0$, $\boldsymbol{A}^2 r_0, \cdots$, $\boldsymbol{A}^{l-1} r_0$ is called Krylov sequence. Note $\mathcal{K}^{l+1}(\boldsymbol{A}, r_0) = \mathcal{K}^l(\boldsymbol{A}, r_0) + \mathrm{span}\left\{\boldsymbol{A}^l r_0\right\}$, and then come to the following conclusion:

(1) $\mathcal{K}^l(\boldsymbol{A}, r_0) \subset \mathcal{K}^{l+1}(\boldsymbol{A}, r_0)$.

(2) $\boldsymbol{A}\mathcal{K}^l(\boldsymbol{A}, r_0) \subset \mathcal{K}^{l+1}(\boldsymbol{A}, r_0)$.

(3) Krylov sequence terminates at the lth time, or there exists l, satisfying $\mathcal{K}^{l+1}(\boldsymbol{A}, \boldsymbol{r}_0) = \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$.

If the orthogonal basis of $\mathcal{K}^l(\boldsymbol{A}, r_0)$ is $\boldsymbol{k}_i$ $(i = 1, 2, \cdots, l)$, denoted $\boldsymbol{K}_l = [\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_l]_{m\times l}$, where $\boldsymbol{k}_i$ is called Lanczos vector (Lanczos, 1950; Desroziers and Berre, 2012).

The best estimates, $\boldsymbol{x}_l \in \mathcal{K}^l(\boldsymbol{A}, r_0)$, of Eq. (A2) satisfies

$$\|\boldsymbol{r}_l\|_2 = \min_{\boldsymbol{x}\in\mathcal{K}^l(\boldsymbol{A}, r_0)} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2, \tag{A3}$$

where $\boldsymbol{r}_l = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l$ is the residual error of $\boldsymbol{x}_l$.

In order to uniquely determine the approximate solution of Eq. (A3), l constraints need to be set. Usually, the residual error is required to satisfy the Petrov-Galerkin condition (Saad, 2003), i.e.,

$$\boldsymbol{r}_l = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l \perp \mathcal{L}^l(\boldsymbol{A}, \boldsymbol{r}_0), \tag{A4}$$

where subspace $\mathcal{L}$ is called constraint space and corresponding subspace $\mathcal{K}$ is called search space. If $\mathcal{L} = \mathcal{K}$, then Eq. (A4) is called the Galerkin condition.

According to the choice of subspace $\mathcal{L}^l(\boldsymbol{A}, \boldsymbol{r}_0)$, different Krylov subspace methods can be used. There are usually two main categories: One is to select $\mathcal{L}^l(\boldsymbol{A}, \boldsymbol{r}_0) = \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$, using Lanczos algorithm (Lanczos, 1950) or CG algorithm; the other is select $\mathcal{L}^l(\boldsymbol{A}, \boldsymbol{r}_0) = \boldsymbol{A}\mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$, using MINRES algorithm.

A2 CG algorithm

For CG algorithm, the constraint space $\mathcal{L}^l(\boldsymbol{A}, \boldsymbol{r}_0) = \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$, so

$$\begin{aligned} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l\|_{\boldsymbol{A}^{-1}} = \min_{\boldsymbol{x}_l\in\mathcal{K}^l(\boldsymbol{A}, r_0)}! &\Leftrightarrow \|\boldsymbol{x}_l - \boldsymbol{x}_*\|_{\boldsymbol{A}} = \min_{\boldsymbol{x}_l\in\mathcal{K}^l(\boldsymbol{A}, r_0)}! \\ &\Leftrightarrow \langle \boldsymbol{x}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l\rangle = 0 \qquad \forall \boldsymbol{x} \in \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0). \end{aligned} \tag{A5}$$

The minimum deviation between the $\|\cdot\|_{\boldsymbol{A}}$ norm of $\boldsymbol{x}_l \in \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$ and the true value $\boldsymbol{x}_* = \boldsymbol{A}^{-1}\boldsymbol{b}$ is equivalent to the $\|\cdot\|_{\boldsymbol{A}^{-1}}$ norm of $\boldsymbol{r}_l = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l$ is minimum, and also equivalent to $\boldsymbol{r}_l$ is orthogonal to $\mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0)$. Where $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle^{1/2}$ and $\|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}} = \left\langle \boldsymbol{x}, \boldsymbol{A}^{-1}\boldsymbol{x}\right\rangle^{1/2}$.

A2.1 *Implementation of Lanczos decomposition*

Equation (A5) represents that the approximate solution $\boldsymbol{x}_l$ found in affine space $\boldsymbol{x}_0 + \mathcal{K}^l$ satisfies:

$$\begin{aligned} &\boldsymbol{x}_l \in \boldsymbol{x}_0 + \mathcal{K}^l \\ &s.t. \qquad \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_l \perp \mathcal{K}^l(\boldsymbol{A}, \boldsymbol{r}_0). \end{aligned} \tag{A6}$$

It is equivalent to

$$\boldsymbol{K}_l^{\mathrm{T}}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_l)=0. \tag{A7}$$

The dimension of Hessian matrix $\boldsymbol{A}$, which is large in the model space, could be reduced by Lanczos decomposition (Saad, 2003):

$$\boldsymbol{A}\boldsymbol{K}_l=\boldsymbol{K}_l\boldsymbol{T}_l+\gamma_l\boldsymbol{k}_{l+1}\boldsymbol{e}_l^{\mathrm{T}}, \tag{A8}$$

where $\boldsymbol{e}_l^{\mathrm{T}}=[0,0,\cdots,1]_l^{\mathrm{T}}$, $\boldsymbol{T}_l$ is a symmetric Hessenberg matrix:

$$\boldsymbol{T}_l=\begin{bmatrix}\delta_1 & \delta_1 & & & 0\\ \gamma_1 & \delta_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{l-2} & \delta_{l-1} & \gamma_{l-1}\\ 0 & & & \gamma_{l-1} & \delta_l\end{bmatrix}. \tag{A9}$$

In $\boldsymbol{T}_l$ matrix, $\delta_l=\boldsymbol{k}_l^{\mathrm{T}}\boldsymbol{A}\boldsymbol{k}_l$, $\gamma_1=\|\boldsymbol{A}\boldsymbol{k}_1-\delta_1\boldsymbol{k}_1\|_2$ and $\gamma_l=\left\|\boldsymbol{A}\boldsymbol{k}_l-\delta_l\boldsymbol{k}_l-\gamma_{l-1}\boldsymbol{k}_{l-1}\right\|_2$.

A2.2 *Solution of Eq. (A7)*

Using Lanczos vectors, the solution of the minimization problem in the l-th iteration can be expressed as:

$$\boldsymbol{x}_l=\boldsymbol{x}_0+\boldsymbol{K}_l\boldsymbol{s}_l. \tag{A10}$$

By substituting Eq. (A10) into Eq. (A7), we can get:

$$\boldsymbol{T}_l\boldsymbol{s}_l=\boldsymbol{K}_l^{\mathrm{T}}r_0, \tag{A11}$$

where $\boldsymbol{s}_l$ is an approximate solution of Eq. (A11). Using the orthogonal basis of $\mathcal{K}^l(\boldsymbol{A},r_0)$, the right side of Eq. (A11) can be expressed as $\boldsymbol{K}_l^{\mathrm{T}}r_0=\rho_0\boldsymbol{e}_l$, where $\boldsymbol{k}_1=\boldsymbol{r}_0/\|\boldsymbol{r}_0\|_2$, $\rho_0=\|\boldsymbol{r}_0\|_2=\sqrt{\boldsymbol{r}_0^{\mathrm{T}}r_0}$, $\boldsymbol{e}_1=[1,0,\cdots,0]^{\mathrm{T}}$.

A2.3 *Recursive relation of CG clgorithms*

Because l is very small, the system of Eq. (A11) can be solved by matrix decomposition, usually using Cholesky or LDL$^{\mathrm{T}}$ decomposition to decompose $\boldsymbol{T}_l$ matrix. The recursive relation of CG algorithm can be obtained as follows:

$$\begin{cases}\alpha_l=\dfrac{\langle\boldsymbol{g}_l,\boldsymbol{g}_l\rangle}{\langle\boldsymbol{p}_l,\boldsymbol{A}\boldsymbol{p}_l\rangle}\\ \boldsymbol{x}_{l+1}=\boldsymbol{x}_l+\alpha_l\boldsymbol{p}_l\\ \boldsymbol{g}_{l+1}=\boldsymbol{g}_l+\alpha_l\boldsymbol{A}\boldsymbol{p}_l\\ \beta_l=\dfrac{\langle\boldsymbol{g}_{l+1},\boldsymbol{g}_{l+1}\rangle}{\langle\boldsymbol{g}_l,\boldsymbol{g}_l\rangle}\\ \boldsymbol{p}_{l+1}=-\boldsymbol{g}_{l+1}+\beta_l\boldsymbol{p}_l\end{cases}, \tag{A12}$$

where α_l is the optimum step size, which makes J minimal in the downward direction, $\boldsymbol{g}_{l+1}$ is the gradient in the new estimation point and $\boldsymbol{p}_{l+1}$ is the new downward direction, which is called search direction.

A3 MINRES algorithm

The constraint space chosen by MINRES algorithm is $\mathcal{L}^l(\boldsymbol{A},r_0)=\boldsymbol{A}\mathcal{K}^l(\boldsymbol{A},r_0)$, and the approximate solution $\boldsymbol{x}_l$ satisfies:

$$\begin{aligned}&\boldsymbol{x}_l\in\boldsymbol{x}_0+\mathcal{K}^l\\ &s.t.\quad \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_l\perp\boldsymbol{A}\mathcal{K}^l(\boldsymbol{A},r_0).\end{aligned} \tag{A13}$$

After l-th iteration, the approximate solution can still be expressed as $\boldsymbol{x}_l=\boldsymbol{x}_0+\boldsymbol{K}_l\boldsymbol{s}_l$. However, unlike CG algorithm directly seek the approximate solution of Eq. (A11), MINRES algorithm seeks the approximate solution $\boldsymbol{s}_l$ to make:

$$\boldsymbol{s}_l=\|\rho_0\boldsymbol{e}_1-\boldsymbol{T}_{l+1,l}\boldsymbol{s}\|_2=\min_{\boldsymbol{s}\in\mathbb{R}^l}!\,. \tag{A14}$$

The $\boldsymbol{T}_{l+1,l}$ matrix here is a $(l+1)\times l$-dimension Hessenberg matrix generated by the Lanczos algorithm:

$$\boldsymbol{T}_{l+1,l}=\begin{bmatrix}\delta_1 & \gamma_1 & & 0\\ \gamma_1 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_{l-1}\\ & & \gamma_{l-1} & \delta_l\\ 0 & & & \gamma_l\end{bmatrix}. \tag{A15}$$

This method avoids the singularity of $\boldsymbol{T}_l$ in Eq. (A11) and avoids the invalidation of the optimal algorithm.

Similar to CG algorithm, because l is very small, QR decomposition is usually used to decompose $\boldsymbol{T}_{l+1,l}$ matrix. The recurrence relation of MINRES algorithm is as follows:

$$\begin{cases} \varepsilon_{l-1} = s_{l-1}\gamma_l \\ \omega_l = c_l c_{l-1}\gamma_l + s_l \delta_{l+1} \\ \tau_{l+1} = \rho_l c_{l+1} \\ \rho_{l+1} = -\rho_l s_{l+1} \\ \theta_{l+1} = c_{l+1}\left(-s_l c_{l-1}\gamma_l + c_l \delta_{l+1}\right) + s_{l+1}\gamma_{l+1} \\ \boldsymbol{p}_{l+1} = \left(\boldsymbol{k}_{l+1} - \varepsilon_{l-1}\boldsymbol{k}_{l-1} - \omega_l \boldsymbol{k}_l\right)/\theta_{l+1} \\ \boldsymbol{x}_{l+1} = \boldsymbol{x}_l + \tau_{l+1}\rho_{l+1} \end{cases}, \tag{A16}$$

where $\boldsymbol{p}_{l+1}$ is the new downward direction, ρ_{l+1} is the new residual error, s_l and c_l are the coefficients of QR decomposition using Givens transform, satisfying $s_l^2 + c_l^2 = 1$.