A conjugate BFGS method for accurate estimation of a posterior error covariance matrix in a linear inverse problem

Yosuke Niwa\textsuperscript{1,2} | Yosuke Fujii\textsuperscript{2,3}

\textsuperscript{1}National Institute for Environmental Studies, Tsukuba, Japan
\textsuperscript{2}Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan
\textsuperscript{3}The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tokyo, Japan

Abstract
One effective data assimilation/inversion method is the four-dimensional variational method (4D-Var). However, it is a non-trivial task for a conventional 4D-Var to estimate a posterior error covariance matrix. This study proposes a method to estimate a posterior error covariance matrix applied to the linear inverse problem of an atmospheric constituent. The method was constructed within a 4D-Var framework using a quasi-Newton method with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The proposed method was constructed such that conjugacy among the set of increment vector pairs was ensured. It is theoretically demonstrated that, when this conjugate property is coupled with preconditioning, an analytical solution of a posterior error covariance matrix could be obtained from the same number of vector pairs as observations. Furthermore, to accelerate the speed of convergence, the method can be coupled with an ensemble approach. By performing a simple advection test, it was confirmed that the proposed method could obtain an analytical matrix of the posterior error covariance within the same number of iterations as the observations. Furthermore, the method was also evaluated using an atmospheric CO\textsubscript{2} inverse problem, which demonstrated its practical utility. The evaluation revealed that the proposed method could provide accurate estimates not only of the diagonal but also of the off-diagonal elements of the posterior error covariance matrix. Although far more expensive than optimal state estimation, the computational efficiency was found to be reasonable for practical use, especially in conjunction with an ensemble approach. The accurate estimation of a posterior error covariance matrix resulting from the proposed method could provide valuable quantitative information regarding the uncertainties of estimated variables as well as the observational impacts, which would be beneficial for designing observation networks. Furthermore, error correlations derived...
1 | INTRODUCTION

Various data assimilation and inversion methods have been used to optimise model variables, such as the initial model states or boundary conditions, in the field of atmospheric sciences. The four-dimensional variational method (4D-Var; e.g. Sasaki, 1969) is an effective data assimilation/inversion method because it can obtain optimal model variables without dealing explicitly with a model’s operator matrix, which is generally too large to store in a computer’s memory. Based on Bayes’ theorem, 4D-Var determines optimal (posterior) model variables through an iterative approach using a prescribed prior estimate and its error covariance, observations, and the error covariance of model–observation mismatches. Ideally, a posterior estimate should have an associated error covariance matrix (a posterior error covariance matrix), although obtaining it is a non-trivial task for conventional 4D-Var approaches.

In this study, we developed a new method to estimate a posterior error covariance matrix within a 4D-Var framework that we apply to an inverse problem of a given atmospheric constituent. In such an inverse problem, we estimate the fluxes of an atmospheric constituent at the Earth’s surface from observations of its atmospheric concentrations. Here, the target tracer is an inert, long-lived tracer such as carbon dioxide (CO₂) and nonlinear chemical processes are not considered. Furthermore, the atmospheric processes considered are all transport-related ones, such as advection, diffusion and convection, and the tracer is treated as a passive tracer. Therefore, the model is linear and results in a linear inverse problem.

Generally, a 4D-Var seeks a set of model variables that minimises a cost function, defined as

\[
J(\mathbf{x}) = \frac{1}{2} (\mathbf{x} - \mathbf{x}^{\text{pri}})^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^{\text{pri}}) + \frac{1}{2} (\mathbf{M}(\mathbf{x}) - \mathbf{y}^o)^T \mathbf{R}^{-1} (\mathbf{M}(\mathbf{x}) - \mathbf{y}^o),
\]

where \( \mathbf{x} \) and \( \mathbf{x}^{\text{pri}} \) are the control vector and its prior estimate, respectively, and \( \mathbf{y}^o \) denotes observations. The sizes of the vectors \( \mathbf{x} \) and \( \mathbf{y}^o \) are \( n \) and \( m \) respectively. \( \mathbf{B} \) and \( \mathbf{R} \) are the error covariance matrices of the prior estimates (\( n \times n \)) and model–observation mismatches (\( m \times m \)), respectively.

\( \mathbf{M}(\cdot) \) is an operator representing a forward model calculation, including an observation operator that transforms variables from the model space into variables in the observation space. In an inverse problem for a given atmospheric constituent, the control vector, \( \mathbf{x} \), and the observation vector, \( \mathbf{y}^o \), generally represent surface fluxes and atmospheric concentrations, respectively, and contain not only spatial but also temporal dimensions for the entire analysis period. Thus, the model operator \( \mathbf{M}(\cdot) \) describes all the model calculations from the beginning to the end. As mentioned, our inverse problem assumes that the model is linear; that is, \( \mathbf{M}(\cdot) \) can be replaced by an \( m \times n \) matrix, \( \mathbf{M} \). Here, the linear model matrix, \( \mathbf{M} \), can be also regarded as the tangent linear operator. Note here that the observation operator included in \( \mathbf{M}(\cdot) \) is also assumed to be linear because we only consider linear spatio-temporal interpolation. Therefore, Equation (1) can be rewritten as

\[
J(\delta \mathbf{x}) = \frac{1}{2} \delta \mathbf{x}^T \mathbf{B}^{-1} \delta \mathbf{x} + \frac{1}{2} (\mathbf{M} \delta \mathbf{x} - \mathbf{d}^o)^T \mathbf{R}^{-1} (\mathbf{M} \delta \mathbf{x} - \mathbf{d}^o),
\]

where \( \delta \mathbf{x} = \mathbf{x} - \mathbf{x}^{\text{pri}} \) and \( \mathbf{d}^o = \mathbf{y}^o - \mathbf{M} \mathbf{x}^{\text{pri}} \). Analytically, the optimal vector, \( \delta \mathbf{x}^{\text{pos}} \), that minimises the cost function (i.e. the posterior estimate) is determined as

\[
\delta \mathbf{x}^{\text{pos}} = \arg \min_{\delta \mathbf{x} \in \mathbb{R}^n} J(\delta \mathbf{x}) = (\mathbf{B}^{-1} + \mathbf{M}^T \mathbf{R}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{R}^{-1} \mathbf{d}^o,
\]

which satisfies

\[
\nabla_{\delta \mathbf{x}} J = \mathbf{B}^{-1} \delta \mathbf{x} + \mathbf{M}^T \mathbf{R}^{-1} (\mathbf{M} \delta \mathbf{x} - \mathbf{d}^o) = 0.
\]

Here, \( \mathbf{M}^T \) represents an adjoint model that performs a backward calculation from concentrations to fluxes. When a cost function is derived from the logarithm of the Gaussian probability density distributions, the posterior error covariance matrix, \( \mathbf{P}_a \ (n \times n) \), is equivalent to the inverse matrix of the Hessian \( \mathbf{A} \). The Hessian is the square matrix of the second-order partial derivatives of the cost function, and in the linear case with the cost function of Equation (2),

\[
\mathbf{A} = \nabla^2 J(\delta \mathbf{x}) = \mathbf{B}^{-1} + \mathbf{M}^T \mathbf{R}^{-1} \mathbf{M}.
\]

Therefore,

\[
\mathbf{P}_a = \mathbf{A}^{-1} = (\mathbf{B}^{-1} + \mathbf{M}^T \mathbf{R}^{-1} \mathbf{M})^{-1}.
\]
In practice, the sizes of the matrices are too large to be stored explicitly in computer memory, rendering the inverse calculation of \((B^{-1} + M^TR^{-1}M)^{-1}\) infeasible. To avoid the calculation of \((B^{-1} + M^TR^{-1}M)^{-1}\), 4D-Var uses either a conjugate gradient method or a quasi-Newton method to search optimal model variables iteratively. Each iterative calculation requires the cost function value and its gradient vector. Therefore, in practice, the posterior error covariance matrix is not calculated explicitly.

Rabier and Courtier (1992) proposed a randomisation method to estimate a posterior error covariance matrix in which an ensemble of perturbed pseudo-observations was employed to approximate the observational part of the Hessian \(M^TR^{-1}M\). Then, the posterior error covariance matrix was obtained by inverting the approximated Hessian matrix using the Sherman–Morrison–Woodbury formula (Sherman and Morrison, 1949). Fisher and Courtier (1995) compared the randomisation method with two other methods: the Lanczos method (Lanczos, 1950) and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Broyden, 1969; Nocedal, 1980; Liu and Nocedal, 1989). The Lanczos method is closely related to a conjugate gradient method, whereas the BFGS method represents one familiar technique of a quasi-Newton method. Therefore, these two methods could provide a posterior error covariance matrix as a by-product of an optimisation. Fisher and Courtier (1995) demonstrated some superiority of the Lanczos method, following comparison of the randomisation, Lanczos and BFGS methods.

Another approach for estimating a posterior error covariance matrix is to use a Monte Carlo method in which the posterior error covariance matrix is estimated from an ensemble spread of posterior estimates. In fact, the idea of estimating an error covariance matrix from an ensemble is the basis of the ensemble Kalman filter (e.g. Evensen, 1994), which is another prominent data assimilation method besides 4D-Var. Within a 4D-Var framework, Chevallier et al. (2007) used the Monte Carlo method to estimate a posterior error covariance applied to a CO₂ inverse problem, in which 4D-Var iterative calculations were performed for each ensemble member. This method was also adopted for another CO₂ inverse system by Liu et al. (2014). However, the Monte Carlo method has trouble estimating off-diagonal elements of a posterior error covariance matrix, because of its stochastic nature. It is well known that stochastic methods usually suffer from sampling errors (e.g. Fisher and Courtier, 1995).

Bousserez et al. (2015), who also targeted a linear inverse problem as we do, proposed a new BFGS method that is combined with the Monte Carlo method (described briefly in Appendix A). In their comparison, the conventional BFGS method was found to be poor at estimating posterior errors (square root of diagonal elements of a posterior error covariance matrix) compared with stochastic methods (i.e. the Monte Carlo and randomisation methods). Through the combination with the Monte Carlo method, their new BFGS method exhibited a performance comparable to stochastic methods for determining posterior errors.

This study proposes a BFGS-based method different from Bousserez et al. (2015), with a particular focus on the linearity of the inverse problem. In general, the quasi-Newton method used with the BFGS algorithm allows some nonlinearity in the optimisation and this is one advantage of the quasi-Newton method. However, our BFGS method has been limited to an exactly linear case. This study, then, leverages a property called “conjugacy.” To our knowledge, the conjugacy has not been focused on conventional studies of atmospheric constituent inversions though the problem is linear. Optimisation methods which exploit conjugacy have been employed to minimise nonlinear cost functions in the context of data assimilation. In particular, several studies (Tshimanga, 2007; Fisher et al., 2008; Tshimanga et al., 2008; Gürol et al., 2014) proposed to use a matrix derived from conjugate vector pairs in preconditioning for accelerating convergence of subsequent assimilation cycles in the so-called incremental 4D-Var (Courtier et al., 1994). In this study, we, however, limit our focus to a linear inverse problem, and we carefully constructed an algorithm that exactly achieves the conjugacy for accurate estimation of a posterior error covariance matrix.

For quantitative evaluation of inversion uncertainties, the estimation of a posterior error covariance matrix requires high accuracy. In an inverse analysis, posterior errors derived from diagonal elements of a posterior error covariance matrix provide valuable information because they would allow for the quantification of the uncertainties of optimised fluxes and the evaluation of the strengths of observational constraints (e.g. Hungershofer et al., 2010; Niwa et al., 2012; Saeki et al., 2013). Furthermore, error correlations derived from off-diagonal elements are also useful for the interpretation of estimated flux variations. An inverse analysis often generates erroneous dipoles in the estimated fluxes due to insufficient observational coverage (e.g. Bousquet et al., 1999). Error correlations enable us to identify such erroneous dipoles.

For an achievement of the high accuracy with a reasonable computation time, we also introduced an ensemble approach as in the method of Bousserez et al. (2015) but achieving the conjugacy. Furthermore, we investigated its feasibility in a realistic inverse problem of an atmospheric constituent and found that the method provides accurate estimation of a posterior error covariance not only for the diagonal elements but also for off-diagonal elements. In this article, we first describe the method starting from the
basic BFGS formula to the techniques introduced for the conjugacy. Then, we evaluate the performance of the proposed method when applied to a simple toy model and a state-of-the-art atmospheric inverse system.

2  |  BASIC METHOD

2.1  |  Quasi-Newton method with the BFGS algorithm

Here, we introduce the principle of the quasi-Newton method combined with the BFGS formula. Readers are encouraged to see Nocedal and Wright (2006) for treatments of the fundamental mathematics and for background theories. Hereafter, we omit δ of the control vector, δx, for simplicity. Also note that, according to conventions, we will use vectors denoted as d and y in subsequent equations to represent a descent direction vector and a gradient increment vector, respectively, which are different from the innovation vector, d₀, and the observation vector, y₀, in the previous equations.

Generally, an optimisation method performs iterative calculations as

\[ x_k = x_{k-1} + \alpha_k d_{k-1}, \]

where \( k \) is the iteration counter, and \( \alpha_k \) is the step length of the line search in the descent direction \( d_{k-1} \) that minimises \( J(x_{k-1} + \alpha_k d_{k-1}) \). Determination of the descent direction \( d_k \) is method dependent. The Newton method determines \( d_k \) using the Hessian A and the gradient of the cost function \( g_k := B^{-1}x_k + M^T R^{-1}(Mx_k - d^o) \) as

\[ d_k = -A^{-1}g_k. \]

If the problem is linear and the cost function is quadratic, then the optimal control vector can be determined through one iteration with \( \alpha = 1 \). This is verified by the fact that Equation (3) is obtained by Equations (6–8). However, in our problem, the matrix size of the Hessian is so large that it renders the Newton method infeasible. Instead, we use a quasi-Newton method that substitutes an approximated inverse Hessian, \( H_k \), as

\[ d_k = -H_k g_k. \]

At each iteration, the approximated inverse Hessian is updated by the BFGS formula as

\[ H_k = (I - \rho_k p_k y_k^T) H_{k-1} (I - \rho_k y_k p_k^T) + \rho_k p_k p_k^T, \]

where \( p_k := x_k - x_{k-1}, \; y_k := g_k - g_{k-1}, \; \rho_k := 1/(y_k^T p_k) \), and \( I \) is the identity matrix \( (n \times n) \). In the above BFGS formula, an arbitrary positive-definite symmetric matrix can be used for the first approximation of the inverse Hessian \( H_0 \). In practice, \( H_k \) is calculated via a two-loop recursion scheme using all the available vector pairs of \( (y, p) \), such that the full matrix of the approximated inverse Hessian is not stored in a computer’s memory (e.g. Nocedal and Wright, 2006).

2.2  |  Estimation of a posterior error covariance matrix with the BFGS formula

Based on the previous algorithm, each column of an approximated inverse Hessian can be obtained by replacing \( g_k \) of Equation (9) with a unit vector such as \( (1, 0, \ldots, 0)^T \), \( (0, 1, \ldots, 0)^T \), \( \ldots \), \( (0, 0, \ldots, 1)^T \). Therefore, if \( N \) pairs of \( (y, p) \) vector pairs are obtained, then the inverse Hessian can be derived via a two-loop recursion scheme (Nocedal, 1980) as follows.

**Algorithm 1** (BFGS-based estimation of the posterior error covariance matrix)

Let \( v \) be the \( j \)th unit vector as

\[ v_j = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \quad \text{for } i = 1, 2, \ldots, n. \]

\[ s \leftarrow v; \]

**for** \( l = N, N - 1, \ldots, 1 \)

\[ \lambda_l \leftarrow \rho_l (p_l^T s); \]

\[ s \leftarrow s - \lambda_l y_l; \]

**end** (for)

\[ s' \leftarrow H_0 s; \]

**for** \( l = 1, 2, \ldots, N \)

\[ \mu \leftarrow \rho_l y_l^T s'; \]

\[ s' \leftarrow s' + (\lambda_l - \mu) p_l; \]

**end** (for)

**stop** with result \( H_N v = s' \), which is the \( j \)th column of \( H_N \).

According to Equation (6), this approximated inverse Hessian, \( H_N \), can be considered the posterior error covariance in the linear case and under the Gaussian assumption. We should note here that \( H_N \) is determined only by the set of the vector pairs of \( (y, p) \) and \( H_0 \). Therefore, the approximation accuracy of \( H_N \) depends on the properties of the \( (y, p) \) vector pairs and \( H_0 \), as well as the available number of \( (y, p) \) pairs. In the next section, we outline the algorithms to prepare \( (y, p) \) vector pairs that have a preferred property of conjugacy. Furthermore, we choose an appropriate positive definite matrix for \( H_0 \). Note here that, in Algorithm 1, the order of \( (y, p) \) vector pairs
does not affect the result. This implies that we do not necessarily need to perform \( N \) iterations of the quasi-Newton method, but simply need to prepare \( N \) vector pairs, which enables us to employ the ensemble approach as described in Section 3.2.2.

3  IMPLEMENTED METHOD

3.1  Theories

3.1.1  Conjugacy

As mentioned in Section 1, our BFGS method is specifically designed for a linear problem, such as the inverse analysis of an atmospheric constituent. Furthermore, our method needs to fulfil the property of mutual conjugacy among the increment vectors; that is,

\[
p_i^T A p_j = 0, \quad \text{for all } i \neq j,
\]

which can be rewritten as

\[
p_i^T y_j = 0, \quad \text{for all } i \neq j.
\]

This conjugacy property presents a favourable condition for the BFGS formula. If the conjugacy is satisfied, then an iterative calculation converges with the analytical solution in at most \( n \) iterations, and it is expected that we can obtain the analytical inverse Hessian after \( n \) iterations (Nocedal and Wright, 2006) as,

\[
H_n = A^{-1}.
\]

Therefore, the conjugacy is expected to enhance the efficiency of the posterior error covariance estimation. Thus, we impose the vector pairs of \((y, p)\) used in the BFGS formula to satisfy Equation (12). To this end, we applied the so-called exact line search and orthogonalization, which are described in Section 3.2.

3.1.2  Preconditioning

Equation (13) indicates that we need \( n \) conjugate vector pairs to obtain an analytical inverse Hessian; however, the necessary number of vector pairs could be further reduced by introducing preconditioning. In an optimisation, preconditioning is often applied to accelerate convergence. One of the most popular preconditioning approaches is that of Lorenc (1988), which transforms the control vector as

\[
\tilde{x} := B^{-1/2}x.
\]

Then, the transformed Hessian can be written as

\[
\tilde{A} = I + B^{1/2} M^T R^{-1} M B^{1/2} = I + XX^T,
\]

where \( X := B^{1/2} M^T R^{-1/2} \). Because \( n > m \), in general, this Hessian has at most \( m \) eigenvalues other than 1 (recall here that \( m \) is the number of observations), which effectively accelerates the convergence speed.

In this study, we also use this preconditioning to estimate a posterior error covariance matrix. By applying the preconditioning to the BFGS method, the transformed approximated inverse Hessian \( \tilde{H}_t = B^{-1/2} H_t B^{-1/2} \) is updated by

\[
\tilde{H}_t = (I - \tilde{p}_i \tilde{p}_i^T) \tilde{H}_{t-1} (I - \tilde{p}_i \tilde{p}_i^T) + \tilde{p}_i \tilde{p}_i^T,
\]

where \( \tilde{p}_i := B^{-1/2} p_i, \tilde{y}_i := B^{1/2} y_i, \tilde{p}_i := 1/(\tilde{p}_i^T \tilde{y}_i) \), for \( i = 1, 2, \ldots, N \). Therefore, \( H_N \) is given with \( \tilde{W}_i := I - \tilde{p}_i \tilde{y}_i \) for \( i = 1, 2, \ldots, N \) as

\[
\tilde{H}_N = \tilde{W}_N^T \cdots \tilde{W}_1^T \tilde{H}_0 \tilde{W}_1 \cdots \tilde{W}_N
\]

\[
+ \tilde{p}_1 \tilde{W}_N^T \cdots \tilde{W}_2^T \tilde{p}_1^T \tilde{W}_2 \cdots \tilde{W}_N
\]

\[
+ \cdots + \tilde{p}_{N-1} \tilde{W}_N \tilde{p}_{N-1}^T \tilde{W}_{N-1} \tilde{W}_N
\]

\[
+ \tilde{p}_N \tilde{W}_N \tilde{p}_N^T.
\]

Because \( \tilde{p}_i = p_i := 1/(\tilde{p}_i^T \tilde{y}_i) \) and \( \tilde{W}_i = B^{1/2} W_i B^{-1/2} \), where \( W_i := I - p_i y_i p_i^T \), the non-transformed approximated inverse Hessian is

\[
H_N = W_N^T \cdots W_1^T \tilde{H}_0 \tilde{W}_1 \cdots W_N
\]

\[
+ \rho_1 W_N^T \cdots W_2^T p_1 p_1^T W_2 \cdots W_N
\]

\[
+ \cdots + \rho_{N-1} W_N^T p_{N-1} p_{N-1}^T W_{N-1} W_N
\]

\[
+ \rho_N W_N p_N^T.
\]

The matrices \( \tilde{H}_0 \) and \( H_0 \) are required to be positive-definite symmetric matrices which approximate \( \tilde{A} \) and \( A \), respectively. It is reasonable to set \( \tilde{H}_0 = I \) due to its simplicity and Equation (15). In this case, the above equation becomes identical to the non-transformed BFGS form with \( H_0 = B \) since \( H_0 = B^{1/2} \tilde{H}_0 B^{1/2} \). Therefore, in practice, by simply setting \( H_0 = B \) in Algorithm 1, we can introduce the preconditioning of Lorenc (1988). Here, it is noted that \( H_0 = B \) is also a reasonable choice because of Equation (5).

When the preconditioning of Lorenc (1988) is applied and the conjugacy is satisfied, the BFGS formula is more favourable than Equation (13); that is, we can obtain the analytical inverse Hessian from only \( m \) \((y, p)\) vector pairs.
that is,
\[ H_m = A^{-1}. \] (19)

This important property can be derived from the fact that \( \hat{y}_k \) and \( \hat{p}_k \) for any \( k \leq m \) can be represented by a linear combination of the same \( m \) orthogonal vectors, when these vector pairs are obtained through iterations of the preconditioned quasi-Newton method, from \( x_0 = 0 \). That is,
\[ \hat{y}_k \in \text{span}\{u_1, u_2, \ldots, u_m\}, \quad \hat{p}_k \in \text{span}\{u_1, u_2, \ldots, u_m\}, \]
for any \( k \leq m \),

where \( \{u_1, u_2, \ldots, u_m\} \) is a set of singular vectors of \( X \). Therefore,
\[ \text{span}\{u_1, u_2, \ldots, u_m\} = \text{span}\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m\}. \] (21)

Detailed proofs of Equations 20 and 21 are given in Appendix B.

Here, we prove Equation (19) as follows. For \( 1 \leq k \leq m \), the conjugacy in the preconditioned space
\[ \hat{p}_i^T \hat{y}_j = \hat{p}_i^T \hat{A} \hat{p}_j = 0, \quad \text{for all } i \neq j, \] (22)
where \( \hat{A} \) is the transformed Hessian, leads to
\[ \hat{H}_k \hat{y}_j = (I - \hat{p}_k^T \hat{y}_j) \hat{H}_{k-1} (I - \hat{p}_k^T \hat{y}_j) + \hat{p}_k^T \hat{p}_k \hat{y}_j \]
\[ = (I - \hat{p}_k^T \hat{y}_j) \hat{H}_{k-1} \hat{y}_j \]
\[ = (I - \hat{p}_k^T \hat{y}_j)(I - \hat{p}_k^T \hat{y}_{k-1}^T) \ldots \]
\[ (I - \hat{p}_j^T \hat{y}_j) \hat{H}_{k-1} \hat{y}_j \] for \( j < k \). (23)

Because
\[ \hat{H}_k \hat{y}_j = (I - \hat{p}_k^T \hat{y}_j) \hat{H}_{j-1} (I - \hat{p}_j^T \hat{y}_j) + \hat{p}_k \hat{p}_j^T \hat{y}_j = \hat{p}_j, \]
for \( j \leq k \) (24)

Equation (23) gives an interesting property (Dennis and Moré, 1977; Fisher and Courtier, 1995) that
\[ \hat{H}_k \hat{y}_j = \hat{p}_j \quad \text{for } j \leq k. \] (25)

Furthermore, an arbitrary vector, \( \bar{u} \), that is orthogonal to every member of \( \{u_1, u_2, \ldots, u_m\} \), satisfies
\[ \hat{H}_k \bar{u} = (I - \hat{p}_k^T \hat{y}_j) \hat{H}_{k-1} (I - \hat{p}_k^T \hat{y}_j) + \hat{p}_k \hat{p}_k^T \bar{u} \]
\[ = (I - \hat{p}_k^T \hat{y}_j) \hat{H}_{k-1} \bar{u} \]
\[ = (I - \hat{p}_k^T \hat{y}_j)(I - \hat{p}_k^T \hat{y}_{k-1}^T) \ldots \]
\[ (I - \hat{p}_1^T \hat{y}_j) \hat{H}_0 \bar{u} = \hat{u}, \] (26)

since \( \hat{p}_k^T \bar{u} = 0 \) and \( \hat{y}_k^T \bar{u} = 0 \) for \( 1 \leq k \leq m \), and \( \hat{H}_0 = I \).

Because of Equation (21), an arbitrary vector, \( \bar{u} \), in the control vector space \( \mathbb{R}^m \) can be written as
\[ \bar{u} = \sum_{i=1}^{m} a_i \hat{p}_i + \bar{u} = \hat{P} \bar{a} + \bar{u}, \] (27)

where \( \hat{P} := (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m) \) and \( \bar{a} := (a_1, a_2, \ldots, a_m)^T \).

Then,
\[ \hat{H}_m \hat{A} \bar{b} = \hat{H}_m (\hat{y} \bar{a} + \bar{u}) = \hat{P} \bar{a} + \bar{u} = \bar{b}, \] (28)

where \( \hat{y} := (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) \), since \( \hat{y} = \hat{A} \hat{p} \), \( \hat{A} = I + B^{1/2} \hat{M} \hat{R}^{-1} \hat{M} B^{1/2} = I + XX^T, \ X\bar{u} = Y \hat{V} \hat{U}^T \bar{u} = 0 \), and Equations (25) and (26). This leads to
\[ (\hat{H}_m \hat{A} - I) \bar{b} = 0. \] (29)

and therefore \( \hat{H}_m \hat{A} = I \), since Equation (29) must be satisfied for an arbitrary vector \( \bar{b} \). Substituting \( \hat{H}_m = B^{-1/2} \hat{H}_m B^{-1/2} \) and \( \hat{A} = B^{1/2} \hat{A} B^{1/2} \), Equation (19) is derived.

In an inversion problem for an atmospheric constituent, the number of observations, \( m \), is usually much smaller than the number of flux parameters, \( n \). Therefore, preconditioning is indispensable given the fact that only \( m \) \( (y, p) \) vector pairs are required to obtain the analytical solution of a posterior error covariance matrix. However, this is only true if the conjugacy is perfectly satisfied.

### 3.2 Preparation of conjugate vector pairs

In order to prepare \( (y, p) \) vector pairs to estimate a posterior error covariance matrix with the BFGS formula, we first perform an optimising calculation. Although the calculation itself is not necessarily the same approach as the estimation of a posterior error covariance, we employ the quasi-Newton method with the BFGS formula herein. To ensure the conjugacy, we introduce the exact line search to an existing quasi-Newton algorithm. Furthermore, we add an ensemble approach to increase the number of \( (y, p) \) vector pairs similarly to Bousserez et al. (2015), but also introducing orthogonalization for the conjugacy.

#### 3.2.1 POPULar with the exact line search

In an optimisation, a line search determines \( a_k \) in Equation (7) such that \( J(x_{k-1} + a_k d_{k-1}) \) is minimised.
Generally, in a nonlinear problem, it requires additional model simulations within iterative quadratic or cubic interpolations to satisfy adequate conditions, such as the so-called Wolfe conditions (e.g., Nocedal and Wright, 2006). In a weakly nonlinear problem, \( \alpha_k = 1 \) is usually used for the initial estimate of \( \alpha_k \) because it is considered adequate and consequently does not require extra simulations at most iterations. However, this is not the exact value that minimises \( f(x_{k-1} + \alpha_k d_{k-1}) \). If a problem is linear, then the cost function is a convex quadratic and the optimal \( \alpha_k \) is analytically determined with the Hessian \( A \) as

\[
\alpha_k = -\frac{d_{k-1}^T g_{k-1}}{d_{k-1}^T A d_{k-1}}. \tag{30}
\]

This is the exact line search. It is known that, if the problem is linear and the exact line search is employed, the quasi-Newton method with the BFGS formula gives conjugate vector pairs (Dennis and Moré, 1977). In fact, that is equivalent to the linear conjugate gradient method (e.g., Nazareth, 1979), which also satisfies the conjugacy (e.g., Nocedal and Wright, 2006).

The practical calculation of Equation (30) is non-trivial because of the huge matrix size of the Hessian \( A \). To avoid dealing explicitly with \( A \), Equation (30) is rewritten as

\[
\alpha_k = -\frac{d_{k-1}^T g_{k-1}}{d_{k-1}^T f_k}, \tag{31}
\]

where

\[
f_k = e_{k-1} + M^T R^{-1} M d_{k-1}. \tag{32}
\]

In Equation (32), we have to perform the adjoint calculation as denoted by \( M^T \); meanwhile we can replace the calculation of \( g_k \) by

\[
g_k = g_{k-1} + \alpha_k f_k, \tag{33}
\]

which no longer includes the adjoint calculation.

The above calculation of the exact line search originates from the method of Derber and Rosati (1989), in which the line search is applied in the linear conjugate gradient method. In this study, we applied this exact line search in the Preconditioned Optimizing Utility for Large-dimensional analyses (POpULar: Fujii and Kamachi, 2003; Fujii, 2005), which uses the quasi-Newton method with the BFGS formula and the preconditioning of Lorenc (1988). One salient feature of POpULar is that it implicitly applies the preconditioning (variables in the preconditioned space, e.g. \( \tilde{x} \) and \( \tilde{g} \), do not appear), while not requiring the calculation of \( B^{-1} \) nor \( B^{1/2} \), which are included in the cost function and gradient calculations, or the variable transformation (Equation (14)). Therefore, it is reasonably flexible for modelling \( B \) and for introducing off-diagonal elements to consider error correlations. The original algorithm of POpULar is designed for nonlinear problems, but in this study, we impose linearity on the algorithm and introduce the exact line search. The original algorithm of POpULar can be found in Fujii (2005) and we only describe the “linear” POpULar as follows.

In the iterative calculation, variables including ancillary variables of \( h := Bg, e := B^{-1}d \) are updated as follows:

**Algorithm 2 (Linear POpULar).**

\[
\begin{align*}
x_0 & \leftarrow 0; \\
g_0 & \leftarrow -M^T R^{-1} d^0; \\
d_0 & \leftarrow -B g_0; \\
e_0 & \leftarrow -g_0; \\
k & \leftarrow 1; \\
\text{repeat} & \\
f_k & = e_{k-1} + M^T R^{-1} M d_{k-1}; \\
x_k & = x_{k-1} + \alpha_k d_{k-1}; \\
& \text{where } \alpha_k \text{ is determined by the exact line search of Equation (31);} \\
g_k & = g_{k-1} + \alpha_k f_k; \\
d_k & = -H_k e_k; \\
e_k & = B^{-1} d_k; \\
k & \leftarrow k + 1; \\
\text{until convergence.}
\end{align*}
\]

In practice, the above updates of \( d_k \) and \( e_k \) are simultaneously calculated using the two-loop recursion scheme as the original POpULar does (Fujii, 2005). It is noted that the original POpULar algorithm replaces the conjugate gradient method of Derber and Rosati (1989) with the quasi-Newton method with a usual (non-exact) line search in order to deal with nonlinear problems. Therefore, the above linear POpULar forms an intermediate between Derber and Rosati (1989) and the original POpULar.

An iterative calculation of the quasi-Newton method often uses the “limited-memory” BFGS algorithm (Nocedal, 1980; Liu and Nocedal, 1989), which uses a limited number of latest \( (y, p) \) vector pairs instead of the full \( k \) pairs. However, we here employ the “full-memory” BFGS algorithm to satisfy the conjugacy among all the vector pairs of \( (y, p) \). In many cases nowadays, as well as an inverse problem that optimises only two-dimensional data of surface fluxes, a 4D-Var calculation can retain all the vectors in storage (but not necessarily in memory) thanks to recent progress in computer technology. Through iterative calculations of the described linear POpULar,
conjugate vector pairs are obtained and they are substituted into the BFGS formula of Algorithm 1 with $H_0 = B$ for estimation of a posterior error covariance matrix.

### 3.2.2 Ensemble with orthogonalization

Bousserez *et al.* (2015) combined the BFGS algorithm with a Monte Carlo method, which functionally increases the number of $(y, p)$ vector pairs. Increasing the number of $(y, p)$ vector pairs according to the ensemble corresponds to an increase in the number of iterations using a single calculation. The total computational cost might not be changed, but the ensemble with application of parallel computation can effectively reduce the wall-clock-time, because each ensemble member can be calculated independently of the others. Here, we use a similar approach to Bousserez *et al.* (2015). However, we newly introduce orthogonalization to ensure the conjugacy among all the vector pairs produced from the ensemble.

To generate the ensemble members, we prepare $M$ sets of $d = (y^0 - Mx^0)$ that are generated simply from random values, where $M$ is the number of ensemble members. Randomly generating $d$ is justified because the posterior error covariance matrix is independent of the observational vector $y^0$ and the prior vector $x^0$ (Equation (6)), and we use the ensemble only to increase the number of $(y, p)$ vector pairs. Thus, we do not need to use actual observations or prior fluxes here. Furthermore, in contrast to the Monte Carlo method of Bousserez *et al.* (2015), it is not required that we perturb the observations according to $R$, nor perturb $x^0$ according to $B$.

After the ensemble calculation, we introduce an orthogonalization procedure that ensures the conjugacy among the vector pairs from all of the ensemble members; within each ensemble member, the $(y, p)$ vector pairs are mutually conjugate, owing to the exact line search. However, when extending the operation beyond a member, this is no longer the case. The conjugacy is not satisfied between different members, because the optimisation itself is independent of other members. In fact, similar orthogonalization is implemented in previous optimisation algorithms as “re-orthogonalization” with the Lanczos method (Fisher, 1998; Gürol *et al.*, 2014). They modify a vector set whose conjugacy should be theoretically achieved but is deteriorated by round-off errors. Those round-off errors are typical of the Lanczos method (Golub and van Loan, 1996).

The orthogonalization is performed as follows. First, we organise all the $p$ and $y$ vectors into rows to form respective matrices

$$P = (p_1, p_2, p_3, \ldots, p_N)$$

and

$$Y = (y_1, y_2, y_3, \ldots, y_N),$$

where $N$ is the total number of $(y, p)$ pairs of all the ensemble members. Then, we multiply $P$ by $Y$ to create a matrix, $Q$, expressed as

$$Q = P^T Y.$$  

Because $Q$ is a symmetric matrix owing to the fact that $p^T y = p^T A p$, and $A$ is symmetric, it can be decomposed as

$$Q = U L U^T,$$

where $U$ and $L$ are the eigenvector and eigenvalue matrices $(N \times N)$ of $Q$, respectively. We should note here that the matrix size of $Q (N \times N)$ is generally not as large compared with that of $H$ or $B (n \times n)$. Theoretically, $Q$ is positive definite and all the eigenvalues are positive. However, too small or negative eigenvalues arise from computational errors. Therefore, those eigenvalues are filtered out from $L$ to ensure computational stability. In this study, we filtered out eigenvalues whose ratio against the maximum eigenvalue was less than $1.0 \times 10^{-6}$. Finally, the $p$ and $y$ vectors are transformed as

$$P' = (p'_1, p'_2, p'_3, \ldots, p'_{N'}) = PU' L'^{-1/2},$$

$$Y' = (y'_1, y'_2, y'_3, \ldots, y'_{N'}) = YU' L'^{-1/2},$$

where $L'$ consists of the effective eigenvalues of $L$, whose size is $N' \times N'$ for $N' < N$, and $U'$ is the corresponding eigenvector matrix, whose size is $N \times N'$. Here, $N'$ is the total number of effective eigenvalues after the filtering. The transformed matrices $P'$ and $Y'$ $(n \times N')$ satisfy

$$P'^{T} Y' = I,$$

and the conjugacy property is satisfied as

$$p'_i^{T} y'_j = p'_i^{T} A p'_j = 0, \quad \text{for all } i \neq j.$$  

To estimate the inverse Hessian in Algorithm 1, these modified vectors of $y'$ and $p'$ and $N'$ replace $y$, $p$, and $N$, respectively. Then, as described in Section 3.1.2, with the preconditioning of $H_0 = B$, the analytical inverse Hessian can be obtained if $N' = m$.

It is theoretically true that the exact line search is not necessarily required when one performs the described orthogonalization, which alone achieves conjugacy among all the vector pairs. However, we recommend the combination of the exact line search and the orthogonalization because we have experienced that their combination stabilises the estimation of the posterior error covariance.
matrix in the experiments that are described below. Furthermore, it is also encouraged because the exact line search does not require any additional computation (the calculation of \( g_k \) that requires the forward and adjoint calculations in a conventional BFGS method is replaced by Equation (32) and \( g_k \) is obtained in the following simple calculation of Equation (33)) and it is sometimes more efficient than the usual line search owing to the fact that an iterative interpolation calculation is not needed to obtain an acceptable value of \( \alpha_k \).

4 | EXPERIMENTS

To evaluate the efficiency and accuracy of our proposed method, we conducted a set of numerical experiments. First, we performed a simple one-dimensional (1D) advection test that was intended to confirm whether the proposed BFGS method worked as theoretically expected. Then, we performed experiments using the case of atmospheric CO2 inversion with a state-of-the-art three-dimensional (3D) model at two different resolutions. The lower-resolution experiment was designed so that we could obtain an analytical matrix of the posterior error covariance to quantify the accuracy of a matrix estimated using the BFGS method. The higher-resolution experiment, in which the numbers of flux parameters were increased in both space and time, was performed to evaluate the practical usability of the method and also to demonstrate values of the posterior error covariance matrix in terms of observational constraint information.

In the evaluations of the aforementioned experiments, we compared our new BFGS method (hereafter denoted as “the conjugate BFGS method”) with another BFGS method proposed by Bousserez et al. (2015), to elucidate the importance of the conjugacy we proposed. Bousserez et al. (2015) used the BFGS formula in conjunction with a Monte Carlo-based ensemble and also introduced cyclic calculations. However, their approach featured a regular quasi-Newton method in which a usual (non-exact) line search was employed, and, thus, the conjugacy is not imposed onto the vector pairs of \((y, p)\). In this study, we modified the cyclic method of Bousserez et al. (2015) (details are described in Appendix A), and combined it with the original POPULAR using a non-exact line search (hereafter denoted as “the non-conjugate BFGS method”). In this method, the conjugacy is unsatisfied, but the same preconditioning of \( H_0 = B \) as in the proposed method is applied. Note that the cyclic method was not used in the conjugate BFGS method because it is not effective when the conjugacy is satisfied (see Appendix A).

4.1 | Simple 1D advection test

For the simple 1D advection test, we made a toy model that advects a Gaussian distribution with a constant wind (10 m s\(^{-1}\)), as shown in Figure 1. The 1D model’s domain ranged from 0 to 1,000 km and the analysis period was 900 min. The interval of the spatial discretization was set as 10 km, that is, \( n = 100 \). Here, the numerical advection was calculated using the van Leer Scheme I (van Leer, 1977) and its adjoint model was also constructed so that the adjoint relationship of \( (M x)^T d_o = x^T (M^T d_o) \) was satisfied. In this experiment, we had five observational locations within the domain (denoted by crosses in Figure 1) and each made three observations during the period with a 300 min interval \( (t = 300, 600, 900 \text{ min}) \); consequently, \( m = 15 \). In this experiment, the parameter to be optimised was the initial state \( (n = 100) \) and its prior estimate was set to zero everywhere in the domain. For simplicity, the error covariance matrices of \( B \) and \( R \) were set as diagonal matrices and their diagonal elements were uniquely set to 0.2 and 0.9, respectively. In this experiment, the ensemble was not used because \( m \) was so small that a single calculation was enough to achieve convergence.

4.2 | CO2 inversion by NICAM-TM 4D-Var

The CO2 inversion experiment used the 4D-Var inverse system, named the NICAM-TM 4D-Var (Niwa et al., 2017a; 2017b), which is based on the Non-hydrostatic ICosa-hedral Atmospheric Model (NICAM: Tomita and
Satoh, 2004; Satoh et al., 2008; 2014)-based Transport Model (NICAM-TM: Niwa et al., 2011). The NICAM adopts an icosahedral grid system consisting of hexagonal and pentagonal grids, instead of the conventional latitude–longitude square grid. The NICAM-TM 4D-Var inverse system uses the forward and adjoint tracer transport models of the NICAM-TM (Niwa et al., 2017a). These models, which are designed “offline”, only calculate tracer transport using prescribed meteorological data. Therefore, they are computationally efficient since the number of iterations required for a 4D-Var calculation is feasible. The forward and adjoint transport models have nonlinear and linear options in their advective calculations. The nonlinearity in the advection calculation stems from having a numerical limiter equipped to achieve tracer monotonicity (Thuburn, 1996; Miura, 2007). In the linear option, that numerical limiter is turned off, achieving the complete linearity of the model as well as the adjoint relationship of $(\mathbf{Mx})^\top \mathbf{d}\mathbf{o} = x^2(\mathbf{M}^\top \mathbf{d}\mathbf{o})$, within the limits of computer machine accuracy; however, monotonicity is not assured and some negative values arise. Here, we used the linear option of the NICAM-TM for the strict linear requirement of the conjugate BFGS method. NICAM-TM 4D-Var originally uses POPULAR for optimisation (Niwa et al., 2017b). In this study, we replaced it with the linear POpULAR as described in Section 3.2.1.

Using NICAM-TM 4D-Var, we performed the low- and high-resolution CO2 inversion experiments adopting the settings of the identical twin experiment by Niwa et al. (2017b). For the low-resolution experiment, we used a reduced horizontal resolution of 480 km (number of grid cells = 2,562) and only the monthly mean flux for July 2010 was optimised. For the high-resolution test, we used a horizontal resolution of 240 km (number of grid cells = 10,242) and the mean fluxes for all 12 months in 2010 were optimised. Therefore, the number of flux parameters to be optimised for the low- and high-resolution experiments were 2,562 (= $n^{\text{low}}$) and 122,904 (= $n^{\text{high}}$), respectively.

For the prior flux error covariance $\mathbf{B}$, we prescribed the diagonal elements based on the prior flux data used in Niwa et al. (2017b). Specifically, a terrestrial flux error (square root of diagonal element of $\mathbf{B}$) represented the square root sum of the squares of the gross primary production (GPP) and respiration (RE) values multiplied by 0.5 (this number was empirically determined). The values of GPP and RE were derived from the terrestrial biosphere model, Vegetation Integrative Simulator for Trace Gases (VISIT: Ito and Inatomi, 2012). The ocean-flux error was defined as the multi-year standard deviation of prior ocean-flux data (Iida et al., 2015). Figure 2a shows the spatial pattern of the prepared prior flux errors for the low-resolution experiment.

To consider error correlations, the off-diagonal elements of $\mathbf{B}$ were defined as

$$B_{ij} = B_i^{1/2} B_j^{1/2} \exp \left( -\frac{l_{ij}^2}{2L_c^2} \right), \quad i \neq j \quad (42)$$

where $l_{ij}$ is the horizontal length between the different flux locations whose element indices are denoted as $i$ and $j$, and $L_c$ is the correlation scale length. Here, $L_c$ was set to 500 and 1,000 km for terrestrial and ocean areas, respectively.

In practice, the full calculation of Equation (42) is computationally expensive, therefore we truncate it by setting $B_{ij} = 0$ for $l_{ij} > 4L_c$. In fact, under a spatially and temporally sparse observation network, introducing off-diagonal elements that consider error correlations can improve posterior flux estimates (e.g. Niwa et al., 2017b).

The synthetic observational data of atmospheric CO2 concentrations assumed weekly flask sampling at 65 ground-based stations (see Niwa et al. (2017b) for the detailed list). Therefore, the numbers of observations for the low- (1-month-long: 4 weeks) and high- (1-year-long: 52 weeks) resolution experiments were 260 (= $m^{\text{low}}$) and 3,380 (= $m^{\text{high}}$), respectively. The geographic locations of the 65 observation sites are shown in Figure 2. For the error covariance matrix of the model–observation mismatches, $\mathbf{R}$, we assumed a diagonal matrix and we set a value of 1 ppm$^2$ for all the diagonal elements.

### 4.3 Diagnostic measures

For the robust evaluation of the accuracy of the estimated posterior error covariance matrices, we used the three diagnostic measures defined in the following items. In the second and third measures below, we denote the estimated and analytically obtained posterior error covariance matrices, the sizes of which are $n \times n$, as $\mathbf{P}^\text{a}_{\text{est}}$ and $\mathbf{P}^\text{a}_{\text{anl}}$, respectively.

- **Total uncertainty**

$$\sigma^\text{tot} = \sqrt{\mathbf{a}^\top \mathbf{P}_a \mathbf{a}}. \quad (43)$$

This measure represents the integrated uncertainty. Vector $\mathbf{a}$ represents the spatial integration operator vector, for example, $\mathbf{a}^\top \mathbf{x}$ represents the global total amount of surface fluxes (scalar) for the global inversion problem. If $\mathbf{x}$ contains different time states, $\mathbf{a}$ can be integrated further in time.
\textbf{F I G U R E 2}  
Distributions of (a) prior and (b) posterior errors of the surface CO$_2$ fluxes in the low-resolution experiment (July). The posterior error was derived through analytical calculation. Magenta triangles denote the locations of the observational sites. In both panels, the errors in ocean areas are too small and thus, they are coloured white.

- Frobenius norm

$$F = \| P_{est}^a - P_{anl}^a \|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (P_{est}^a_{ij} - P_{anl}^a_{ij})^2}. \quad (44)$$

This quantifies the difference between $P_{est}^a$ and $P_{anl}^a$ using the square root of the sum of the squared differences of all the elements.

- Divergence

$$D = \frac{1}{2n} \left( \text{tr}\left[ P_{anl}^a (P_{est}^a)^{-1} \right] + \text{tr}\left[ (P_{est}^a)^{-1} P_{anl}^a \right] \right) - 1. \quad (45)$$

This measure also represents a difference but in another aspect; it was first introduced and named divergence by Ueno and Nakamura (2016). Divergence is non-negative and it becomes zero if $P_{est}^a = P_{anl}^a$.

In the simple 1D advection test and the low-resolution CO$_2$ inversion experiment, we can easily obtain $P_{anl}^a$ because of its small matrix size ($n = 100$ and 2,562 for the simple 1D advection test and the low-resolution inversion experiment, respectively). We obtained each $P_{anl}^a$ by analytically calculating Equation (6), by performing forward simulations with a pulse flux at each model grid cell and their corresponding adjoint simulations $n$ times, followed by the inverse calculation of the $n \times n$ matrix. In contrast, for the high-resolution CO$_2$ inversion experiment, $P_{anl}^a$ was not available due to its huge matrix size. Therefore, we only used the measure of the total uncertainty (Equation (43)). Figure 2b shows the distribution of the posterior flux errors derived using the analytical calculation in the low-resolution CO$_2$ inversion experiment. The posterior flux error is defined as the square root of a diagonal element of the posterior error covariance matrix, that is, $\sigma_1 := \sqrt{P_{anl}^a_{ii}}$.

\section{RESULTS}

\subsection{Simple 1D advection test}

In Figure 3, we show the results of the simple advection test using the conjugate and non-conjugate BFGS methods. As a reference, we show the cost function changes along iterations in the parameter (initial state) optimisation in Figure 3a, in which the results by the exact and non-exact line searches, that is, by the linear POpULar and the original POpULar, respectively, are presented. Note here that the calculation of the parameter optimisation is independent of the estimation of the posterior error covariance matrix. Figure 3a shows that this experiment was so simple that the convergence of the optimisation was achieved within just two iterations, and this was true regardless of whether the exact line search was used.

However, in the estimation of the posterior error covariance matrix, the convergence required many more iterations and the exactness of the line search, that is, conjugacy became important. Figure 3b–d show the diagnostic measures of the estimated posterior error covariance matrices that were derived from the ($y, p$) vector pairs of the optimisation calculations. At least 15 iterations were required for convergence with the conjugate BFGS method. This number is consistent with the number of iterations theoretically required by the preconditioning
FIGURE 3 Changes of the diagnostic variables along with the number of iterations, in the simple 1D advection test: (a) cost function, (b) total uncertainty, (c) Frobenius norm, and (d) divergence. The black line with open circles indicates the results from the conjugate Broyden–Fletcher–Goldfarb–Shanno (BFGS) method proposed in this study and the grey line with crosses depicts the results by the non-conjugate BFGS method (Appendix A).

(\(m\), i.e. 15 in this case, see Section 3.1.2). However, with the non-conjugate BFGS method, convergence was not achieved, even after 15 iterations. Furthermore, the calculation was stopped before any convergence due to its computational instability (not shown). After 15 iterations, the conjugate BFGS method achieved almost zero values for the Frobenius norm and the divergence, and reached 6.3 for the total uncertainty; this was the same as the analytically obtained value, indicating that the conjugate BFGS method worked as theoretically expected. In the non-conjugate BFGS method, the cyclic method was used with the non-exact line search and it improved the performance to some extent (Appendix A), but never reaching the level of the conjugate BFGS method. This fact demonstrates the significant advantage of the exact line search used in the conjugate BFGS method. Note here that the ensemble was not used given the small size of the problem.

Figure 4 presents distributions of the posterior error and how the estimated distribution changed with number of iterations, compared to the analytically obtained distribution. At the 15th iteration, the conjugate BFGS method produced an error distribution that is no longer distinguishable from the analytical one. However, the non-conjugate BFGS method could not reproduce the analytical distribution well.

The off-diagonal elements of the estimated posterior error covariance matrix were also investigated using a posterior error correlation that was calculated as

\[
    r_{ij} = \frac{p_{ij}}{p_i^{1/2}p_j^{1/2}}, \quad i \neq j. \tag{46}
\]

Figure 5 presents the posterior error correlations at the 15th iteration derived by the conjugate and non-conjugate BFGS methods, compared with the ones that were analytically obtained. The error correlations estimated by the conjugate BFGS method were coincident with the analytical set, again demonstrating the theoretical realisation of the conjugate BFGS method. However, the error correlations were poorly calculated using the non-conjugate BFGS method, and erroneous correlations were produced for remote locations.

5.2 Low-resolution CO2 inversion experiment

As demonstrated above, just replacing the usual non-exact line search by the exact one achieved the pronounced improvements for the estimation of the posterior error covariance matrix. This manifests the importance of the conjugacy we proposed. In this section, we show the results of the CO2 inversion experiments, in which the
number of observations was much larger and therefore we required the introduction of the ensemble.

5.2.1 | Single case

First, we show how the methods performed without the ensemble, that is, using only one member as in the previous simple 1D advection test. Figure 6a presents the cost function reduction associated with increased numbers of iterations. It shows that the exactness of the line search does not largely affect the convergence speed in the optimisation of the control variables (i.e. the flux estimation), which is the same behaviour as in the simple 1D advection test (here, the optimisation used the linear and original POPULar algorithms for the conjugate and non-conjugate cases, respectively). In both cases, approximately 15 iterations are required for convergence.

Figure 6b–d show how closely the estimated matrix matches the analytical one, as the number of iterations increased. As in the simple 1D advection test, the posterior error covariance estimation requires more iterations for convergence than the flux estimation. Furthermore, the convergence speed of the conjugate BFGS method is much faster than that of the non-conjugate one, as was the case in the simple 1D advection test. Nevertheless, the conjugate BFGS method did not fully achieve convergence within 100 iterations. The estimated total uncertainty almost reached the true value of 0.32 PgC year\(^{-1}\) after 100 iterations and the divergence was also nearly zero after 70 iterations; however, the Frobenius norm did not achieve convergence, even after 100 iterations. This result also indicates that the Frobenius norm is the strictest measure of the three.

Figure 7 shows the relative “error” distributions of the posterior errors estimated using the conjugate BFGS method with 10, 50 and 100 iterations, and that from the non-conjugate BFGS method with 100 iterations; these values are represented as percentages of the analytical values, that is, \(\frac{\sigma_{\text{est}}^2 - \sigma_{\text{anl}}^2}{\sigma_{\text{anl}}^2} \times 100\%\). Figure 7 demonstrates that, in the conjugate BFGS method, the posterior errors were insufficiently estimated during the first stage, especially in regions with dense observations (e.g. Europe), but that such insufficiency was gradually mitigated as the iteration proceeded. After 100 iterations, the posterior errors were nearly compatible with the analytical ones. The estimate of the off-diagonal elements was similarly improved (Figure 8). Here, we evaluated the off-diagonal elements using the error correlations defined by Equation (45), and \(j\) of Equation (45) was fixed to a grid-point flux on East Asia; the distributions of \(r_{ij}\) along \(i\) are presented. Figure 8 shows that the distribution of the error correlation gradually approached the analytical results in the conjugate BFGS method. However, both the Figures 7 and 8 show that the non-conjugate BFGS method did not sufficiently estimate the posterior error covariance, even after 100 iterations.
FIGURE 7 Accuracy of the posterior errors estimated by the conjugate BFGS method at the (a) 10th, (b) 50th, and (c) 100th iterations, and (d) that by the non-conjugate BFGS method at the 100th iteration, which are all derived from a single member in the low-resolution CO₂ inversion. Also shown are those for the 50-ensemble member cases with 15 iterations: (e) conjugate, (f) non-conjugate. Here, the error is defined as \[ \frac{\sigma_{\text{est}}^i - \sigma_{\text{true}}^i}{\sigma_{\text{true}}^i} \times 100 \], where \( \sigma_{\text{est}}^i = \sqrt{P_{\text{est}}^i} \) and \( \sigma_{\text{true}}^i = \sqrt{P_{\text{true}}^i} \) respectively. Errors <2% are coloured white.

As mentioned earlier, the conjugate BFGS method did not fully achieve convergence even at the 100th iteration. This is not surprising because convergence theoretically requires 260 iterations (= \( m^{\text{low}} \)), at most, as described in Section 3.1.2. Nevertheless, the state of the model nearly approached convergence at the 100th iteration, indicating that the number of effective eigenvalues of the second term of the preconditioned Hessian (Equation (15)) was much smaller than 260. However, this slow convergence might not be suitable for practical use, although it depends on the degree of accuracy required. Therefore, we adopt the ensemble approach to accelerate convergence.

5.2.2 Ensemble case

Figure 9 shows how the use of the ensemble accelerated the convergence speed. Here, we used 50 ensemble members that were arbitrarily generated. As described in Section 3.2.2, all the vector pairs of \( (y, p) \) in the conjugate BFGS method were modified to achieve the conjugacy, using the orthogonalization in conjunction with the exact line search. We also performed the calculation using the non-conjugate BFGS method with 50 ensemble members.

Figure 9 demonstrates that, when the 50 ensemble members are used in the conjugate BFGS method, the
convergence achieves almost the same or a much faster speed than that of the parameter optimisation (flux estimation) shown by Figure 6a. The total uncertainty reached an analytical value of 0.32 PgC·month$^{-1}$ after 13 iterations and the divergence dropped to zero within a few iterations. Even with the strictest measure, that is, the Frobenius norm, convergence was achieved within approximately 15 iterations and the value reached near zero. As in the single-member case, this convergence speed was much faster than that of the non-conjugate BFGS method. The

**FIGURE 8** Distributions of the posterior error correlations against the flux in the grid cell centred on 28° N, 115° E, derived using the conjugate BFGS method at the (a) 10th, (b) 50th, and (c) 100th iterations, and (d) that by the non-conjugate BFGS method at the 100th iteration; these are all derived from a single member in the low-resolution CO$_2$ inversion. Also shown are those values for the 50-ensemble member cases with 15 iterations ((e) conjugate, (f) non-conjugate) and (g) the analytical calculation.
non-conjugate BFGS method required several times more iterations than the conjugate one; the total uncertainty reached the analytical value after 31 iterations and the convergence of the Frobenius norm was achieved after approximately 50 iterations. The effectiveness of the conjugate BFGS method can also be demonstrated by the distributions shown in Figures 7e and 8e, which were derived using only 15 iterations but were identical to the analytical ones. Meanwhile, this is not the case for the non-conjugate BFGS method, as shown by Figures 7f and 8f.

With the conjugate BFGS method, we investigated the sensitivity to the number of ensemble members. When the number of the ensemble members was reduced from 50 to 10, the required number of iterations for convergence increased threefold; the total uncertainty reached the analytical value at the 39th iteration and the Frobenius norm approximately converged after 60 iterations. This demonstrates that we can obtain a faster convergence speed with a larger number of ensemble members. However, the iteration numbers required for convergence in the 10-member case are not five times larger than those of the 50-member case, suggesting that the efficiency of the ensemble decreases as the number of members employed increases.

As described in Section 3.2.2, in the conjugate BFGS method, all the \((y, p)\) pairs were transformed to \((y', p')\) for conjugacy, and some vector pairs were filtered out as part of this transformation. Consequently, the available number of \((y', p')\), \(N'\), must have been smaller than the original available number of \((y, p)\), \(N\). Figure 10a shows the total uncertainty change with \(N'\) in the 10- and 50-member cases. It clearly demonstrates that the convergence speed against \(N'\) is almost the same for the 10- and 50-member cases, indicating that \(N'\) determines the estimation accuracy of the posterior error covariance matrix. Figure 10b shows the change in the ratio \(N'/N\) associated with the number of iterations. It indicates that the transformation filters out more vectors as the iteration proceeds, and that its extent is enhanced with larger ensemble members, which in fact induced the decrease of the ensemble efficiency. In summary, these results suggest that the matrix, \(Q\), of Equation (36) produced from more iterations would be better conditioned than that produced by a larger ensemble, if \(N\) were the same. In practice, if the total computational costs were limited, it would be better to perform a large number of iterations with a small number of ensemble members. In contrast, if one wanted to prioritise computational speed and if massive parallel computing were feasible, a large number of ensemble members might be preferable.

5.3 High-resolution CO₂ inversion experiment

Finally, we investigate the feasibility of the conjugate BFGS method in a practical problem and also demonstrate the utility of the posterior error covariance matrix. In this
experiment, we used a typical spatio-temporal parameter size that adopted the half-grid size of the previous low-resolution experiment, that is, approximately 240 km, and we estimated 12 monthly mean fluxes for the entire year instead of only 1 for the month of July.

In this experiment, approximately 20 iterations were required for the convergence of the flux optimisation (Figure 11a). As already demonstrated, the estimation of the posterior error covariance matrix required many more iterations than the flux estimation, and the same was true even when 50 ensemble members were employed (Figure 11b). With 50 members, the conjugate BFGS method dramatically reduced the total uncertainty within a few iterations, although this reduction subsequently slowed. With a larger ensemble size of 200, the convergence feature was similar but the total uncertainty was significantly smaller than with the 50 members.

When one sees the number of \((y', p')\) vector sets, \(N'\), the convergence appears to require \(N' > 1,000\) (Figure 11c), which corresponds to 46 iterations using the 200 ensemble members. This suggests the required value of \(N'\) is only 10 times larger than that of the low-resolution experiment (Figure 10a). This number is similar to the increased ratio of the number of the observations, that is, \(m_{\text{high}}/m_{\text{low}} = 13\) (from 4 to 52 weeks), rather than that of the flux parameters, that is, \(n_{\text{high}}/n_{\text{low}} = 48\) (4 times in space and 12 times in time). In fact, this is consistent with the theoretical fact that the number of observations determines the number of iterations required for convergence; at most \(m\) iterations are required if the conjugacy is satisfied and the preconditioning of Lorenc (1988) is used (Section 3.1.2).

Thus, although a larger number of ensemble members are required than in the low-resolution experiment, we confirmed the feasibility of the developed method at the high resolution, which is practically applied using real observations. A posterior error covariance matrix estimated for a 1-year period would provide useful information that varies seasonally. One valuable diagnostic parameter that can be derived from the posterior error covariance matrix is the error reduction ratio, which is defined as

\[
 r_i = \frac{\sigma_{\text{pos}}^i - \sigma_{\text{pri}}^i}{\sigma_{\text{pri}}^i} \times 100\%, \tag{47}
\]

where \(\sigma_{\text{pos}}^i\) and \(\sigma_{\text{pri}}^i\) are posterior and prior errors, respectively, and \(i\) denotes each flux component. This explains how errors have decreased from their corresponding prior errors and it is useful to diagnose where, when, and by how much the observations constrain flux estimates, in the inverse analysis. Monthly distributions of the error reduction ratio defined by Equation (47) are shown in Figure 12.

In the series of the distributions, we find clear temporal variation in the observational impacts over terrestrial areas. In boreal summer (May–October), the error reduction ratios are notable, whereas they are marginal in other months. This is especially true for inland areas located far from the observation sites. This is attributable to the larger prior errors seen in terrestrial areas in the summer, which have enough room to be constrained by the observations. In contrast, in the winter, the prior errors of terrestrial areas are comparatively small, resulting in strong prior constraints. Furthermore, the stratified atmosphere of winter might have limited the observational impacts near the observation sites. Ocean fluxes are tightly constrained by small prior errors and, consequently, the observational impacts are generally limited.

Figure 13 shows the posterior error correlations against the flux in East Asia for July and January, which are similar to those in Figure 8e, but instead are derived from the high-resolution experiment and focus on Asian regions. Compared with Figure 8e, the spatial pattern for July looks generally similar but shows more detailed structures in East Asia (Figure 13a). Surrounding areas around target flux have positive correlations. In fact, those correlations...
were inherited from the prior error covariance, which was defined by Equation (42) using $L_c = 500$ km. Beyond these positive correlations, terrestrial grids exhibit negative correlations, followed by positive correlations in part, just like a ripple. In January, the positive correlations surrounding the target flux become pronounced, indicating a larger contribution from the prior error correlation. This is consistent with the smaller error reduction ratios shown by Figure 12, that is, weaker observational constraints in January than in July for this region. Furthermore, negative correlations show a different spatial pattern in July, located on the western side of the positive correlations. These posterior error correlation distributions are determined by the prior errors, airflow patterns, model–observation mismatch errors, and timings and locations of observations. The former two factors differ by seasons, causing such different correlation distributions.

To evaluate the global features of the posterior error correlations, we shrank the error covariance matrix through aggregation as

$$ P_{agg} = GP_aG^T, \quad (48) $$

where $G$ is an operator matrix that spatially integrates fluxes by regions and $P_{agg}$ is the aggregated posterior error covariance matrix. Figure 14 shows the posterior errors and their correlations among aggregated regions for each month, which were derived from diagonal and off-diagonal elements of $P_{agg}$, respectively. Here, the aggregation was performed by integrating the fluxes into 13 areas comprising 11 terrestrial regions (A–K) defined by the TransCom 3 project (Gurney et al., 2002; Baker et al., 2006), the global ocean (L) and the global total (M). In Figure 14, we found a number of existing negative
correlations at each month. This indicates that, even after the aggregation, the fluxes were not independently estimated. Specifically, error correlations persistently exist throughout the year between tropical America (C) and temperate South America (D), between northern (E) and southern Africa (F) and between tropical (I) and temperate Asia (H), thus indicating that it is difficult to clearly distinguish fluxes within the continents of South America, Africa and southern Asia. Meanwhile, some correlations differ according to season. We found significant negative correlations between boreal (A) and temperate North America (B) for May–September, and between boreal (G) and temperate Asia (H) for July; they are not pronounced in the other seasons.

Plotting error correlations against the global total (M) are also useful, which are depicted in the bottom row of each panel in Figure 14. They are positive values and represent the size of each regional flux estimate’s contribution to the global total uncertainty. Figure 14 shows that significant positive correlations against the global total uncertainty persisted in the regions of South America (C and D), Africa (E and F) and tropical Asia (I). They were in fact caused by significant observational gaps in these regions. Such positive correlations quantify the observational insufficiency for the global CO₂ flux estimate. This would help us to interpret the information contained in the global flux estimates and also to consider the future design of global observational networks.

6 | DISCUSSION

This study proposed a BFGS-based method for estimating a posterior error covariance matrix that is novel in its focus on the conjugacy among \( (y, p) \) vector pairs for a linear inverse problem of an atmospheric constituent; it achieves a high accuracy and efficient (theoretically expected) convergence speed. In fact, the basic idea of our method, using the conjugate vector pairs in the BFGS formula, is the same as “the quasi-Newton Limited-Memory
Preconditioner (QN-LMP)” referred to by Tshimanga et al. (2008). However, differently from the study of Tshimanga et al. (2008), which aims at acceleration of optimisation, we devoted the method to estimate a posterior error covariance matrix as accurate as possible for elaborating uncertainties of an atmospheric constituent inversion.

According to Tshimanga (2007), if the information of all the previous iterations is exploited, QN-LMP is analytically equivalent to “the Ritz LMP,” which is recommended by Tshimanga et al. (2008). The Ritz-LMP method uses approximated eigenvalues and eigenvectors, which are so-called Ritz pairs (Golub and van Loan, 1996). The drawback of QN-LMP is that it requires twice as much computer memory as Ritz-LMP. This is true for our method, because it employs two series of the vectors \((y, p)\). However, it is compatible with the ensemble approach that we additionally introduced, because the inevitable non-conjugacy between different members can be modified by the orthogonalization with the simple multiplication of \(p\) by \(y\) (Equation (36)). In addition to the requirement of the two series of the vectors, we found that the method needs a few hundred ensemble members for practical use, which is much larger than that of Bousseretz et al. (2015) (a few members). Furthermore, we fully store the information obtained from all the previous iterations (i.e. not “limited-memory”). However, storing all these data is feasible in a practical inverse analysis of an atmospheric constituent, because the control vector consists of surface fluxes (two-dimensional in space).

Fisher and Courtier (1995) and Meirink et al. (2008) used eigenvalues and eigenvectors of the Hessian obtained by the combination of the conjugate gradient method and the Lanczos method and they calculated the posterior error covariance matrix with an equation equivalent to “the spectral Limited-Memory Preconditioner LMP” referred to by Tshimanga et al. (2008). In practice, their eigenvalues and eigenvectors are approximated by Ritz pairs from a limited number of iterations and are not exact ones. If the Ritz-LMP is applied, their method would provide an equivalent posterior error covariance matrix to that from our method. It is also noted that, in the Lanczos method, the orthogonality among the approximated eigenvectors tends to deteriorate as the number of the vectors increases due to computational errors (Golub and van Loan, 1996).

One limitation of our method is that it imposes linearity on a model. This linearity should be ensured in a discretised programme level. In this application, using a model that only deals with linear processes (e.g. atmospheric transport) is not enough because such a model often includes a nonlinear numerical scheme in discretised programmes, intentionally added for a given reason, such as ensuring tracer monotonicity. Any such nonlinear numerical scheme should be turned off in advance to ensure compatibility with our BFGS-based method. Furthermore, its corresponding adjoint model is required so that the adjoint relationship is satisfied.

Application of the conjugate BFGS method to a nonlinear problem (e.g. with a nonlinear numerical scheme) would not work well without modification. First, the exact line search is not applicable to a nonlinear problem. This is readily understood from the result of Niwa et al. (2017b) that the linear conjugate gradient method, which uses the exact line search and hence is equivalent to the conjugate BFGS method, failed to optimise fluxes when applied to a nonlinear model. In addition, the orthogonalization procedure for the conjugacy is also not applicable because \(Q\) \((= P^TY)\) becomes asymmetric when the problem is nonlinear. Modification of the conjugate BFGS method for nonlinear problems is, thus, a future study.

Nevertheless, it may well be that a nonlinear (well-behaved) numerical scheme improves model accuracy and consequently provides better estimates in an inverse analysis. In fact, Niwa et al. (2017b) found this to be true for NICAM-TM 4D-Var, which has both linear and nonlinear options. Meanwhile, Niwa et al. (2017b) also found that the nonlinear effect is limited to a small degree, indicating that the cost function structure of the nonlinear case, and hence its Hessian, might not differ largely from that of the linear case. Therefore, in an intrinsically linear inverse problem, it would be acceptable to, on the one hand, use a linear model for posterior error covariance estimation and, on the other hand, use a nonlinear model for flux estimation.

The most prominent feature of our method is the highly accurate (almost analytical) estimation with a reasonably feasible computational cost. As shown in Figure 9, the non-conjugate BFGS method can achieve almost the same accuracy as our method, but it requires several times more iterations. Furthermore, the non-conjugate BFGS method requires a greater computational cost when calculating the BFGS formula. Our method first performs inexpensive orthogonalization and reduces the number of \((y, p)\) vector pairs from \(N\) to \(N'\). For instance, the reduction ratio was about 90% at the 60th iteration in the low-resolution \(CO_2\) inversion experiment, with 50 members (Figure 10). However, the non-conjugate BFGS method fully uses \(N\) vector pairs, and when the cyclic method is employed, the practical number of the vector pairs introduced in the BFGS formula is \(L \times N\), where \(L\) is the number of cycles. Therefore, the calculation of the BFGS formula differs by two or more orders between the conjugate and non-conjugate methods. This is critical when the number of model parameters, \(n\), is large, and in
fact, it is computationally demanding for the typical size of an inverse analysis.

Another virtue of our method is its flexibility regarding the initial ensemble perturbation. Because of its stochastic nature, the Monte Carlo method has to perform the perturbation along with $\mathbf{B}$ and $\mathbf{R}$, which often requires eigenvalue decomposition of those matrices (Chevallier et al., 2007). However, our method generates ensemble members using plain random numbers, and therefore, it does not require the decomposition of any large-size matrix. Although eigenvalue decomposition should be performed in Equation (37), its matrix size is much smaller than $\mathbf{B}$. Furthermore, we coupled the conjugate BFGS method with the optimisation scheme named POPULAR, in which the implicit preconditioning is employed to avoid the calculations of $\mathbf{B}^{-1}$ and $\mathbf{B}^{1/2}$ and the exact lines search is implemented without any additional numerical cost. Thus, the required computational techniques are completely simple. The preconditioning is imperative for efficient estimation of a posterior error covariance matrix. As described in Section 3.1.2, the preconditioning assures convergence within at most the number of observations, which is generally smaller than the number of the parameters in an inverse problem of an atmospheric constituent.

Nevertheless, the method’s computational cost should be further reduced in the future because it is more expensive than the flux estimation. A wall-clock-time can be reduced by introducing an ensemble with parallel computing but its efficiency decreases as the number of members increases, as shown by Figure 10b. One prospective way to improve the ensemble efficiency might lie in the initial ensemble perturbation; however, the convergence speed was not significantly altered by the different initial perturbations we have tested so far (changing perturbation magnitudes or perturbing along with $\mathbf{B}$ and $\mathbf{R}$, as the Monte Carlo method does). Further investigations and developments might be required but they are left for future study.

7 CONCLUSIONS

This study proposed to use a method based on the BFGS formula for estimating a posterior error covariance matrix in the inversion of an atmospheric constituent. By limiting the method to an exactly linear case in use and ensuring the conjugacy, we achieved a high accuracy and efficiency. The method can use the ensemble technique to accelerate the convergence speed and hence, it is compatible with massive parallel computing. Our method employed the exact line search to satisfy the conjugacy among the available $(\mathbf{y}, \mathbf{p})$ vector pair set. When using the ensemble approach, further transformation of the $(\mathbf{y}, \mathbf{p})$ vectors through eigenvalue decomposition, ensured conjugacy beyond an ensemble member. We proved that the conjugacy combined with the preconditioning yields a preferable property: only the same number of vector pairs as the observations are required for it to converge with the analytical solution of a posterior error covariance matrix. This fact was confirmed by conducting a simple 1D advection test. The test clearly demonstrated the importance of the conjugacy for efficient convergence. The experiment of the CO$_2$ inverse problem showed that the method has the ability to produce a satisfactory accurate posterior error covariance matrix within a reasonable number of iterations, not only for diagonal but also for off-diagonal elements. Furthermore, compared with another ensemble that does not achieve the conjugacy, our method is superior in terms of convergence speed. Such accurate estimation of a posterior error covariance matrix would be useful for quantitative evaluation of observational impacts and for design of optimal observational networks. Furthermore, it could also yield insights regarding posterior flux estimates by elucidating their independency.

ACKNOWLEDGEMENTS

This study was supported mainly by the Environment Research and Technology Development Fund (2-1401 and 2-1701) of the Ministry of the Environment, Japan (JPMEERF20142001 and JPMEERF20172001). The work was also supported by the JSPS KAKENHI Grant Number 19K03976, the ISM Cooperative Research Program (2019-ISMCRP-2030) and the FLAGSHIP2020, MEXT, within priority study 4 (advancement of meteorological and global environmental predictions utilising observational “Big Data”). Y. N. thanks the staff members at JAMSTEC, RIKEN, and the University of Tokyo for developing the NICAM. The calculations of this study were performed on the FUJITSU PRIMEHPC FX100 supercomputer system of the Meteorological Research Institute.

ORCID

Yosuke Niwa  https://orcid.org/0000-0002-7600-9816

REFERENCES

Baker, D.F., Law, R.M., Gurney, K.R., Rayner, P., Peylin, P., Denning, A.S., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fung, I.Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S. and Zhu, Z. (2006) TransCom 3 inversion intercomparison: impact of transport model errors on the interannual variability of regional CO$_2$ fluxes,
1988–2003. *Global Biogeochemical Cycles*, 20(1). https://doi.org/10.1029/2004GB002439.

Bousquet, P., Ciais, P., Peylin, P., Ramonet, M. and Monfray, P. (1999) Inverse modeling of annual atmospheric CO₂ sources and sinks: 1. Method and control inversion. *Journal of Geophysical Research*, 104(D21), 26161–26178.

Bousserez, N., Henze, D.K., Perkins, A., Bowman, K.W., Lee, M., Liu, J., Deng, F. and Jones, D.B.A. (2015) Improved analysis-error covariance matrix for high-dimensional variational inversions: application to source estimation using a 3D atmospheric transport model. *Quarterly Journal of the Royal Meteorological Society*, 141(690), 1906–1921.

Broyden, C. (1969) A new double-rank minimisation algorithm. Preliminary report. *American Mathematical Society, Notices*, 16(4), 670.

Chevallier, F., Bréon, F.-M. and Rayner, P.J. (2007) Contribution of the Orbiting Carbon Observatory to the estimation of CO₂ sources and sinks: theoretical study in a variational data assimilation framework. *Journal of Geophysical Research*, 112, D09307. https://doi.org/10.1029/2006JD007375.

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519), 1367–1388.

Dennis Jr, J.E. and Moré, J.J. (1977) Quasi-Newton methods, motivation and theory. *SIAM Review*, 19, 46–89.

Derber, J. and Rosati, A. (1989) A global oceanic data assimilation system. *Journal of Physical Oceanography*, 19, 1333–1347.

Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5), 10143–10162.

Fisher, M. and Courtier, P. (1995) *Estimating the covariance matrices of analysis and forecast error in variational data assimilation*. Technical Memorandum 220. Reading, UK: ECMWF.

Fisher, M., Nocedal, J., Trémolet, Y. and Wright, S.J. (2008) Data assimilation in weather forecasting: a case study in PDE-constrained optimization. Technical Memorandum 561. Reading, UK: ECMWF.

Fujii, Y. (2005) Preconditioned optimizing utility for large-dimensional analyses (POpULar). *Journal of Oceanography*, 61, 167–181.

Fujii, Y. and Kamachi, M. (2003) A nonlinear preconditioned quasi-Newton method without inversion of a first-guess covariance matrix in variational analyses. *Tellus A*, 55, 450–454.

Golub, G. H. and Van Loan C. F. (1996) *Matrix Computations*, 3rd edition. Baltimore and London, MD: Johns Hopkins University Press.

Gurney, K.R., Law, R.M., Denning, A.S., Rayner, P.J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., Fung, I.Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Maki, T., Maksyutov, S., Masarie, K., Peylin, P., Prather, M., Pak, B.C., Randerson, J., Sarmiento, J., Taguchi, S., Takahashi, T. and Yuen, C.-W. (2002) Towards robust regional estimates of CO₂ sources and sinks using atmospheric transport models. *Nature*, 415, 626–630.

Gürol, S., Weaver, A.T., Moore, A.M., Piacentini, A., Arango, H.G. and Graton, S. (2014) B-preconditioned minimization algorithms for variational data assimilation with the dual formulation. *Quarterly Journal of the Royal Meteorological Society*, 140(679), 539–556.

Hungrershoefker, K., Breon, F.-M., Peylin, P., Chevallier, F., Rayner, P., Klonecki, A., Houweling, S. and Marshall, J. (2010) Evaluation of various observing systems for the global monitoring of CO₂ surface fluxes. *Atmospheric Chemistry and Physics*, 10, 10503–10520.

Iida, Y., Kojima, A., Takatani, Y., Nakano, T., Sugimoto, H., Midorikawa, T. and Ishii, M. (2015) Trends in pCO₂ and sea–air CO₂ flux over the global open oceans for the last two decades. *Journal of Oceanography*, 71, 637–661. https://doi.org/10.1007/s10772-015-0306-4.

Ito, A. and Inatomi, M. (2012) Use of a process-based model for assessing the methane budgets of global terrestrial ecosystems and evaluation of uncertainty. *Biogeoosciences*, 9, 759–773.

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, 255–282.

van Leer, B. (1977) Towards the ultimate conservative difference scheme. IV. A new approach to numerical conservation. *Journal of Computational Physics*, 23, 276–299.

Liu, J., Bowman, K.W., Lee, M.M., Henze, D.K., Bousserez, N., Brix, H., Collatz, G. J., Menemelis, D., Ott, L., Pawson, S., Jones, D. and Nassar, R. (2014) Carbon monitoring system flux estimation and attribution: impact of ACOS-GOSAT XCO₂ sampling on the inference of terrestrial biospheric sources and sinks. *Tellus B*, 66(1), 22486. https://doi.org/10.3402/tellusb.v66.22486.

Liu, D.C. and Nocedal, J. (1989) On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45, 503–528.

Lorenc, A.C. (1988) Optimal nonlinear objective analysis. *Quarterly Journal of the Royal Meteorological Society*, 114(479), 205–240.

Meirink, J.F., Bergamaschi, P. and Krol, M.C. (2008) Four-dimensional variational data assimilation for inverse modelling of atmospheric methane emissions: method and comparison with synthesis inversion, *Atmos. Chem. Phys.*, 8, 6341–6353. https://doi.org/10.5194/acp-8-6341-2008.

Miura, H. (2007) An upwind-biased conservative advection scheme for spherical hexagonal–pentagonal grids. *Monthly Weather Review*, 135, 4038–4044.

Nazareth, L. (1979) A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms. *SIAM Journal on Numerical Analysis*, 16, 794–800.

Niwa, Y., Fujii, Y., Sawa, Y., Iida, Y., Ito, A., Satoh, M., Imasu, R., Tsuboi, K., Matsueda, H. and Saigusa, N. (2017b) A 4D-Var inversion system based on the icosaheiral grid model (NICAM-TM 4D-Var v1.0). Part 2: Optimization scheme and identical twin experiment of atmospheric CO₂ inversion. *Geoscientific Model Development*, 10, 2201–2219.

Niwa, Y., Machida, T., Sawa, Y., Matsueda, H., Schuck, T.J., Brenninkmeijer, C.A.M., Imasu, R. and Satoh, M. (2012) Imposing strong constraints on tropical terrestrial CO₂ fluxes using passenger aircraft based measurements. *Journal of Geophysical Research*, 117, D11303. https://doi.org/10.1029/2012JD017474.

Niwa, Y., Tomita, H., Satoh, M. and Imasu, R. (2011) A three-dimensional icosaheiral grid advection scheme preserving monotonicity and consistency with continuity for atmospheric tracer transport. *Journal of the Meteorological Society of Japan*, 89, 255–268.

Niwa, Y., Tomita, H., Satoh, M., Imasu, R., Sawa, Y., Tsuboi, K., Matsueda, H., Machida, T., Saskawa, M., Belan, B. and Saigusa, N. (2017a) A 4D-Var inversion system based on the icosaheiral grid
model (NICAM-TM 4D-Var v1.0). Part 1: Offline forward and adjoint transport models. *Geoscientific Model Development*, 10, 1157–1174.

Nocedal, J. (1980) Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35, 773–782.

Nocedal, J. and Wright, S.J. (2006) *Numerical Optimization*, 2nd edition. New York, NY: Springer.

Rabier, F. and Courtier, P. (1992) Four-dimensional assimilation in the presence of baroclinic instability. *Quarterly Journal of the Royal Meteorological Society*, 118(506), 649–672.

Satoh, M., Tomita, H., Yashiro, H., Miura, H., Kodama, C., Seiki, R. J. and Belikov, D. (2013) Carbon flux estimation for Siberia by inverse modeling constrained by aircraft and tower CO2 measurements. *Journal of Geophysical Research – Atmospheres*, 118, 1100–1122. doi:https://doi.org/10.1002/jgrd.50127.

Satoh, M., Matsuno, T., Tomita, H., Nasuno, T. and Iga, S. (2014) The non-hydrostatic icosahedral atmospheric model (NICAM-TM 4D-Var v1.0). Part 1: Offline forward and dynamic and theoretical, innumerical variational objective analysis. *Journal of the Royal Meteorological Society*, 134(632), 751–769.

Sherman, J. and Morrison, W.J. (1949) Adjustment of an inverse matrix corresponding to changes in the elements of a given row or a given column of the original matrix. *Annals of Mathematical Statistics*, 20, 620–624.

Thuburn, J. (1996) Multidimensional flux-limited advection schemes. *Journal of Computational Physics*, 123, 74–83.

Tomita, H. and Satoh, M. (2004) A new dynamical framework of nonhydrostatic global model using the icosahedral grid. *Fluid Dynamics Research*, 34, 357–400.

Tshimanga, J. (2006) On a class of limited memory preconditioners for large-scale nonlinear least-squares problems (with application to variational ocean data assimilation). PhD thesis, University of Namur, Belgium, 167 pp.

Tshimanga, J., Gratton, S., Weaver, A.T. and Sartenaer, A. (2008) Limited-memory preconditioners with application to incremental four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 134(632), 751–769.

Ueno, G. and Nakamura, N. (2016) Bayesian estimation of the observation-error covariance matrix in ensemble-based filters. *Quarterly Journal of the Royal Meteorological Society*, 142(698), 2055–2080.

**How to cite this article:** Niwa Y, Fujii Y. A conjugate BFGS method for accurate estimation of a posterior error covariance matrix in a linear inverse problem. *Q J R Meteorol Soc*. 2020;146:3118–3143. https://doi.org/10.1002/qj.3838

**APPENDIX A. THE NON-CONJUGATE BFGS METHOD (THE MODIFIED BOUSSEREZ’S METHOD)**

In this study, we modified the method of Bousserez et al. (2015) and have referred to this form as “the non-conjugate BFGS method” in the evaluation of our conjugate BFGS method. Bousserez et al. (2015) similarly used the BFGS formula, but they proposed using a diagonal matrix derived from the diagonal elements of the approximated inverse Hessian at the previous iteration, that is,

\[
H_0^k = \text{diag}(H^{k-1}),
\]  

(A1)

where \( k \) is the iteration counter. Note here that, in practice, one may use a different initial approximation of the inverse Hessian at each iteration if it is a positive definite symmetric matrix. It is expected that Equation (A1) improves the initial approximation of the inverse Hessian as the iteration proceeds. Furthermore, Bousserez et al. (2015) also introduced a method that performs the BFGS calculation cyclically, whereby the same \((y, p)\) vector pairs are used several times while the initial approximation of the inverse Hessian is updated using the previous cycle as Equation (A1). From this cyclic method, the initial approximation of the inverse Hessian is further improved as the cycle proceeds, resulting in an improved estimate of a posterior error covariance matrix.

Here, we adopt the approach of the cyclic method in the BFGS calculation of Algorithm 1, but modify the initial approximation of the inverse Hessian. Instead of Equation (A1), we simply use the full matrix of the approximated inverse Hessian of the previous cycle as

\[
H_0^l = H^{l-1},
\]  

(A2)

where \( l \) represents the counter of the cycles and changes from 1 to \( L \). This is achieved by simply reusing the \((y, p)\) vector sets \( L \) times, which uses almost the same equation as Equation (10):

\[
H_i = (I - \rho p_i y_i^T)H_{i-1}(I - \rho y_i p_i^T) + \rho p_i p_i^T,
\]  

(A3)

where \( i \) ranges from 1 to \( M(=L \times N) \). Note here that \( N \) is the available number of \((y, p)\) vector pairs. When an ensemble is used, \( N \) is the total number of \((y, p)\) pairs from all the ensemble members. The above \((y, p)\) vectors satisfy

\[
y_i = y_j \text{ and } p_i = p_j \text{ for } i = j \text{ (mod } N) \text{ and } 1 \leq j \leq N.
\]  

(A4)

This modification is intended to retain the off-diagonal elements of the approximated inverse Hessian matrix. In the method of Bousserez et al. (2015), Equation (A1) resets
the off-diagonal elements to zero at each cycle, which might constrict the improvement of the accuracy of the initial approximation of the inverse Hessian, when the off-diagonal elements are significantly large. This is especially true when a prior error covariance matrix is designed as a non-diagonal matrix. For the initial approximation in Equation (A3), we set \( H_0 = B \) as the conjugate BFGS method proposed in this study.

Note here that the method we proposed does not employ this cyclic calculation. That is because if the conjugacy is satisfied, then adding the same vector pairs into the BFGS formula does not make a difference. This is proven later. In this method, the \((y, p)\) vector pairs are prepared similarly through iterative calculations of the BFGS-based quasi-Newton method as the conjugate BFGS method, but with a usual (non-exact) line search; that is, we used the original POpULAr (Fuji, 2005).

Bousserez et al. (2015) combined their BFGS method with a Monte Carlo method. We also coupled the above modified method with an ensemble approach, but not with the Monte-Carlo method. When an ensemble is used, we perturb observations around \( M(x^{\text{pos}}) \), where \( x^{\text{pos}} \) is the posterior flux vector obtained by a single optimisation calculation, that is,

\[
y^{\alpha,i} = M(x^{\text{pos}}) + d^{\alpha,i},
\]

where \( i \) is an ensemble member index and \( d^{\alpha,i} \) is the perturbation vector that is generated by simple random values (the same as our proposed method). Bousserez et al. (2015) perturbed observations according to \( R \) and also perturbed \( x^{\text{ori}} \) according to \( B \) as well. However, we found that this observation-only perturbation was enough, because there was almost no difference between the above simple perturbation and the Monte Carlo perturbation in the CO2 inversion experiment (not shown).

In the analysis that compared our proposed method with the non-conjugate BFGS method (described in the main text), we coherently set \( L = 10 \). One may increase this number although a larger number requires a proportionately more computational cost. Figures A1 and A2 show how this cyclic calculation improved the results.

Furthermore, we here prove that the cyclic method does not affect the estimate of a posterior error covariance matrix when all the \((y, p)\) vector pairs are conjugate to each other. First, we rewrite the BFGS formula of Equation (A3) as

\[
H_M = W_M^T \cdots W_1^T H_0 W_1 \cdots W_M + \rho_1 W_M^T \cdots W_1^T p_i^0 p_j^T W_2 \cdots W_M \\
\ldots \\
+ \rho_{M-1} W_M^T p_{M-1}^0 p_{M-2}^T W_{M-1} \cdots W_M
\]

If the conjugate property of \( p_i^T y_j = 0 \) for all \( i \neq j \) but \( i = j \) (mod \( N \)) is satisfied, \( W_i \) and \( W_j \) are commutative as

\[
W_i W_j = I - \rho y_i^T p_j^T - \rho y_j^T p_i^T = W_j W_i. \tag{A7}
\]

Furthermore, a multiple product of \( W_i \) equals \( W_i \) as

\[
W_i W_i = I - 2 \rho y_i^T p_i^T + \rho^2 y_i^T y_i^T p_i^T = I - 2 \rho y_i^T p_i^T + \rho y_i^T p_i^T = I - \rho y_i^T p_i^T = W_i, \tag{A8}
\]

since \( \rho_i = 1/y_i^T p_i \). Besides,

\[
p_i^T W_j = p_i^T \text{ for all } i \neq j \text{ but } i = j \text{ (mod } N) \tag{A9}
\]

and

\[
p_i^T W_j = p_i^T - \rho p_i^T y_j^T p_j^T = 0, \text{ for } i = j \text{ (mod } N) \tag{A10}
\]

are also satisfied. Using Equations (A7–A10), Equation (A6) can be rewritten as

\[
H_M = W_N^T \cdots W_1^T H_0 W_1 \cdots W_N + \rho_1 p_1^0 p_1^T + \rho_2 p_2^0 p_2^T \cdots + \rho_N p_N^0 p_N^T. \tag{A11}
\]

We also obtain Equation (A11) from Equation (A6) but \( M \) is replaced by \( N \); thus, \( H_M = H_N \). Therefore, the cyclic method does not provide any additional information about the posterior error covariance matrix.

**APPENDIX B. PROOF OF EQUATIONS 20 AND 21 UNDER THE PRECONDITIONING AND THE CONJUGACY**

Now, the preconditioned gradient is given as

\[
\tilde{g} = \tilde{x} + B^{1/2} M^T R^{-1} (MB^{1/2} \tilde{x} - d^0). \tag{B1}
\]

If we set \( \tilde{x}_0 = 0 \) in the preconditioned quasi-Newton method, the initial gradient is given as

\[
\tilde{g}_0 = X \delta,
\]

where \( \delta := - R^{-1/2} d^0 \). Furthermore, the \( n \times m \) matrix of \( X \) is decomposed as

\[
X = U \Sigma V^T, \tag{B3}
\]

where \( U \) and \( V \) are \( n \times n \) and \( m \times m \) orthogonal matrices, respectively, and their columns \((u_1, u_2, \ldots u_n \text{ and } v_1, v_2, \ldots v_m)\) are the left- and right-singular vectors of \( X \), respectively; and \( \Sigma \) is an \( n \times m \) rectangular diagonal matrix,
FIGURE A1  Plots of the three diagnostic measures of the estimated posterior error covariance matrix: (a) total uncertainty, (b) Frobenius norm, and (c) divergence in the low-resolution CO$_2$ inversion. The panels show the results from a single member with the non-conjugate BFGS method without the cyclic calculation (solid line) and with 10 cycles (dotted line) (same as “non-conjugate” in Figure 6b–d).

FIGURE A2  Same as Figure 9, but the grey solid line indicates the non-conjugate BFGS method without the cyclic calculation. The grey dotted line is the same as Figure 9 (the non-conjugate BFGS method with 10 cycles). Both the two lines are derived from 50 ensemble members whose diagonal elements ($\sigma_1, \sigma_2, \ldots, \sigma_m$) are singular values of $X$. Then, $\delta$ is represented by a linear combination of the right-singular vectors as

$$\delta = \sum_{i=1}^{m} a_i v_i.$$  

Therefore, $\tilde{g}_0$ can be rewritten as

$$\tilde{g}_0 = \sum_{i=1}^{m} b_{0,i} u_i,$$

where $b_{0,i} = \sigma_i a_i$.  

This gradient is used as the first search direction, consequently,

$$\tilde{p}_1 = \tilde{x}_1 = -\alpha_1 \tilde{g}_0 = -\sum_{i=1}^{m} c_{1,i} u_i,$$  

where $c_{1,i} = \alpha_1 b_{0,i}$.  

Here, let us make the induction hypothesis that $\tilde{x}_{k'}, \tilde{p}_{k'}$ and $\tilde{g}_{k'}$ are expressed as linear combinations of $\{u_1, u_2, \ldots, u_m\}$ for $k' \leq k$. Then, $\tilde{g}_k$ can be also expressed as a linear combination of $\{u_1, u_2, \ldots, u_m\}$ as follows:

$$\tilde{x}_k = \sum_{i=1}^{m} c_{k,i} u_i,$$

(B6)

$$\tilde{g}_k = \sum_{i=1}^{m} c_{k,i} u_i + X [X^T \left( \sum_{i=1}^{m} c_{k,i} u_i \right) + \delta]$$

(B7)

$$= \sum_{i=1}^{m} c'_{k,i} u_i + X \sum_{i=1}^{m} a_i v_i = \sum_{i=1}^{m} b_{k,i} u_i,$$

where $c'_{k,i} = (1 + \sigma_i^2) c_{k,i}$ and $b_{k,i} = c'_{k,i} + \sigma_i a_i$. Consequently, $\tilde{y}_k (= \tilde{g}_k - \tilde{g}_{k-1})$ is also expressed as a linear combination of $\{u_1, u_2, \ldots, u_m\}$. Moreover, $\tilde{d}_k$ could be a linear combination of $\{u_1, u_2, \ldots, u_m\}$, if $\tilde{d}_k$ is derived by $\tilde{g}_k$, $\tilde{p}_k$, and $\tilde{y}_k$ that are all expressed as linear combinations of $\{u_1, u_2, \ldots, u_m\}$ for $1 \leq k' \leq k$. For instance, by the preconditioned BFGS formula, the next search direction $\tilde{d}_k$ can be derived by Equation (17) as

$$\tilde{d}_k = \tilde{W}_k^T \cdots \tilde{W}_1^T \tilde{W}_0 \tilde{W}_1 \cdots \tilde{W}_k \tilde{g}_k,$$
It is obvious that, in the right-hand side, all but the first term are represented by linear combinations of \{u_1, u_2, \ldots, u_m\}. Furthermore, the first term is also a linear combination of \{u_1, u_2, \ldots, u_m\}, because $H_0 = B$, and hence $\tilde{H} = I$. Therefore, $\tilde{x}_{k+1} = \tilde{x}_k + \alpha_k \tilde{d}_k$ and $\tilde{p}_{k+1} = \tilde{x}_{k+1} - \tilde{x}_k$ are also linear combinations of \{u_1, u_2, \ldots, u_m\}. Thus, the induction has proven that all $\tilde{y}_k$ and $\tilde{p}_k$ for $1 \leq k \leq m$ are linear combinations of \{u_1, u_2, \ldots, u_m\}. Since $p$ is linearly independent when the conjugacy is satisfied, the aforementioned property leads to an important fact that

$$\text{span}\{u_1, u_2, \ldots, u_m\} = \text{span}\{\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m\}. \quad (B9)$$