Framework for the comparison of \textit{a priori} and \textit{a posteriori} error variance estimation and tuning schemes

Michael Sitwell | Richard Ménard

Abstract
The performance of an assimilation system is strongly dependent on the quality of the error statistics used. A number of error statistics estimation and tuning methods have previously been developed to better assess and determine these statistics. Many of these are \textit{a posteriori} methods which make use of quantities calculated during the assimilation procedure, while other \textit{a priori} methods do not require information from the assimilation. In this study, we develop a conceptual framework that relates these methods when applied to error variance determination, where each method is associated with the minimization of a particular cost function. The minimization of these cost functions describes a fitting procedure that fits parts of the prescribed modelled innovation covariance to its observed values. Each method must in some way separate the innovation covariance into its contributions from the background and the observations, which are then used in the fitting procedure. It is shown that the examined \textit{a posteriori} methods use the analysis filter to make this separation and that the minimization of their associated cost functions is done implicitly within the tuning procedure. Analytical expressions for the expectation value and variance of estimates for error variance scaling parameters are determined for each method. The expressions for the expectation values of these estimates show that the accuracy of each method is dependent on its ability to separate the background from the observation contributions to the innovation covariance. This separability is quantified by use of the Frobenius inner product between the background- and observation-error covariances, which additionally allows for geometric interpretations of the covariances to be made. Comparisons between variance parameter estimates from different methods are made for the case of a 1D periodic domain.

KEYWORDS
data assimilation, Desroziers diagnostic, Hollingsworth–Lönnberg method, innovation statistics, error tuning, estimation of error statistics
1 | INTRODUCTION

Data assimilation is the process of statistically blending information from observations with a model and is widely used in atmospheric sciences, particularly for numerical weather forecasting. The degree of influence the observations have on the blended model state, known as the analysis, is dictated by the error covariances assigned to the observations and prior model state in the assimilation process. The analysis increment (the quantity added to the background state to form the analysis) is strongly influenced by the background-error covariance that describes the prior model uncertainty; it spatially spreads out information contained in the observations and, for multivariate assimilation, can impose balance between different variables (Bannister 2008 gives a review). As the analysis increment lies in the subspace spanned by the background-error covariance matrix (Bannister, 2008), accurately specifying the background-error covariance is of particular importance in data assimilation. Background-error covariances are influenced by, for example, the growth of errors in the forecasting model and, in an assimilation cycle, are also influenced by past observation accuracies and their spatial and temporal distributions. In addition to instrument uncertainties, the observation-error covariances include representativeness errors and errors from imperfect observation operators (Janjić et al., 2018). The performance of an assimilation system depends on how accurately these error covariances are specified (Daley, 1992).

In practice, obtaining covariances that accurately describe these uncertainties can be difficult and many simplifying assumptions are often made (Rutherford, 1972; Parrish and Derber, 1992; Evensen, 1994; Polavarapu et al., 2005). As such, methods to diagnose and tune error statistics are of significant interest. Methods that use the statistics of assimilation residuals fall into two basic types: a priori diagnostic methods that use observation-minus-background residuals (known as innovations when the observations are assimilated) and a posteriori diagnostics that require analyses.

The method of Hollingsworth and Lönnberg (1986) (HL; actually first introduced in Rutherford 1972) is a well-known a priori method that consists of calculating the autocorrelation statistics of observation-minus-background residuals as a function of the separation distance between a pair of observation sites. In this method, assuming that the observation errors are spatially uncorrelated (or have a small length-scale in comparison to the spatial correlations of the background errors) allows one to distinguish and estimate the statistics of the observation and background errors. The HL method has been successfully extended to estimate multichannel observation-error covariances (Garand et al., 2007; Bormann and Bauer, 2010; Bormann et al., 2010), temporal correlations of observation errors (Macpherson and Laroche, 2019), and to incorporate spatial observation-error correlations (Eresmaa and Järvinen, 2005). In the context of atmospheric chemistry, chemical error statistics have been obtained by using the HL method with a single polar-orbiting satellite by using the distance between consecutive measurements along the satellite track (Ménard et al., 2019).

The maximum likelihood method (Dee and Da Silva, 1999) is another a priori method that involves the minimization of a cost function in observation space. This method requires knowledge of the probability distribution of the innovations (usually assumed to be normally distributed). While the maximum likelihood method can easily become computationally expensive, it has been used in meteorology (Dee et al., 1999; Tangborn et al., 2002) and in air quality (Ménard et al., 2016). The maximum likelihood method has been used to estimate error variances, background and observation correlation length-scales, and cross-covariances between observation and background errors (Tangborn et al., 2002). A priori error statistics are also used in consistency diagnostics, such as the \( \chi^2 \) diagnostic (Bennett, 1992; Ménard and Chang, 2000) and the variance and covariance consistency diagnostics (Ménard, 2016).

The development of a posteriori methods began with the so-called \( J_{\text{min}} \) consistency diagnostic introduced in Tarantola (1987) and Talagrand (1999). Desroziers and Ivanov (2001) (DI01) adapted this method to iteratively estimate covariance scaling parameters for both the background and observation covariances in a variational assimilation system. Chapnik et al. (2004) extended the DI01 method to diagnose scaling parameters for subsets of the error covariances. Chapnik et al. (2004) also showed that the DI01 method is equivalent to maximum likelihood estimation for normally distributed innovations. The DI01 method was tested with an operational system in Chapnik et al. (2006), but the computational cost of carrying out the estimations remained high.

Another a posteriori method was developed by Desroziers et al. (2005) (D05) based on statistical properties of analysis residuals. As the D05 method is often easier to apply for high-dimensional problems, this method gained popularity in numerical weather prediction, especially for the estimation of satellite observation-error covariances (Bormann et al., 2010; Stewart et al., 2014; Weston et al., 2014; Todling, 2015; Gauthier et al., 2018). The D05 method has also been used to estimate model-error variances in ensemble Kalman filtering (Li et al., 2009), Doppler radar radial wind error correlations (Waller et al., 2016b), spatial and temporal observation-error
correlations for ground-based Global Navigation Satellite System (GNSS) Zenith Total Delay observations (Bennitt et al., 2017; Macpherson and Laroche, 2019), and observation errors in multi-species atmospheric chemical constituent assimilations (Skachko et al., 2016). Although devised as an iterative algorithm, most operational applications of the D05 algorithm only apply one iteration due to the added computational cost of computing multiple analyses.

Properties of the D05 algorithm were first derived in the context of a simplified observation network by Desroziers et al. (2005). Waller et al. (2016a) examined how mis-specifications of the initial error variances and correlation lengths influences the D05 estimates of observation errors in a simplified observation network. Ménard (2016) considered the iterative estimation of both the observation- and background-error covariances in a simplified observing network and showed that convergence to the true error covariances in the D05 scheme is possible if the error correlations are correctly specified. Ménard (2016) also established that to estimate the observation-error covariances alone requires that the background-error covariance be correctly specified, which was later shown by Bathmann (2018) to be a general conclusion and not restricted to the simplified observation network. Similarly, Gauthier et al. (2018) concluded that the iterative estimation of the observation-error covariance using the D05 method requires a good estimation of the background-error covariance.

While a number of the studies referenced above have examined the behaviour of the D05 algorithm, there have been relatively few comparisons between the D05, DI01, and HL methods. In one of the few studies that compare the D05 and HL methods, Macpherson and Laroche (2019) found that the two methods produced significantly different results for the temporal correlations for GNSS Zenith Total Delay observation errors. The lack of a mathematical formalism for these comparisons contributes to the difficulty of comparing these methods. One of the objectives of this study will be to develop this formalism. This unifying framework will allow for more straightforward comparisons between algorithms. Furthermore, we will develop a geometrical interpretation of the error covariances which will allow us to interpret these algorithms geometrically as well as provide a simple set of descriptive quantities that summarize the error statistics used in assimilation. We will also extend these interpretations to consistency diagnostics such as the $\chi^2$ diagnostic. Beyond the new framework developed to compare methods, the statistical properties of the algorithms will be presented and compared to one another. In this work, we restrict our analysis to the estimation of error variances for simplicity. Accordingly, this assumes that the error correlations have already been adequately specified.

We note in passing that methods other than those discussed above have been developed to estimate forecast-error statistics in model space, such as Peña and Toth (2014) and Feng et al. (2020), and to estimate forecast sensitivities to error statistics parameters such as Langland and Baker (2004), Daescu and Todling (2010), Daescu and Langland (2013), and Hotta et al. (2017).

This paper is organized as follows. A brief review of assimilation is given in Section 2. Section 3 discusses the role of filtering in assimilation and Section 4 introduces the Frobenius inner product, which will be used frequently in this paper. Sections 5, 6, and 7 examines the DI01, D05, and HL methods, respectively, and cast each method into a form that allows for a direct comparison between methods. The geometric interpretation of covariances will be detailed in Section 8, while some of the statistical properties of each algorithm are examined in Section 9. Results of the DI01, D05, and HL methods for the case of homogeneous covariances defined on a 1D periodic domain are discussed in Section 10. Section 11 further discusses the theoretical differences between the DI01 and D05 algorithms and conclusions are given in Section 12.

2 | PRELIMINARIES

Before examining assimilation error statistics more closely, we give a brief review of the analysis procedure. The impact the observations have on the analysis is set by specifying a background-error covariance $\tilde{B}$ and an observation-error covariance $\tilde{R}$, which in general differ from their true values $B$ and $R$, respectively. Here and throughout this paper, prescribed (or modelled) quantities used in assimilation are denoted with a tilde (‘’) while their true values are denoted with the superscript ($\uparrow$).

Observation-minus-background residuals (innovations) are often key quantities when assessing the error covariances used in assimilation. The innovation vector $d$ is given by $d = y - H(x_b)$, where $y$ is the vector of observations, $x_b$ is the background (or trial) model field, and $H$ is the (in general) nonlinear observation operator. The innovation vector can be approximated to linear order as $d \approx e_o - He_b$, where $e_o$ and $e_b$ are the observation and background errors, respectively, and $H$ is the linearized observation operator. In this approximation, the innovation involves both the background and observation errors without reference to the true state of either variable, but only provides a sum of these errors and not each error individually. If there is no correlation between the background and observation errors (Rutherford, 1972; Daley,
of data assimilation), so that
\[ S^T \bar{R} S = \Phi. \] (3b)

The diagonal elements of \( \Phi \) then represent the ratio of the observation to background-error variances in the simultaneously diagonalized bases. The full description of constructing the operator \( S \) is described in Appendix A. Looking at this transformation in another light, this transformation solves the generalized eigenvalue problem \( \bar{R} S = (\bar{H} \bar{B} \bar{H}^T) \Phi S \), so that \( S \) and \( \Phi \) are matrices of the generalized eigenvectors and eigenvalues, respectively. Note that this decomposition is general and only assumes that \( \bar{H} \bar{B} \bar{H}^T \) and \( \bar{R} \) are positive definite matrices and does not assume that they represent homogeneous or isotropic covariances, or that the observations are uniformly spaced. In the special case where \( \bar{H} \bar{B} \bar{H}^T \) and \( \bar{R} \) are defined on a uniformly spaced periodic 1D domain, the simultaneous eigenvectors are the Fourier modes. In the general case, the simultaneous eigenvalues of \( \bar{H} \bar{B} \bar{H}^T \) and \( \bar{R} \) are referred to as the “spectra” of these matrices, even when this decomposition does not coincide with the Fourier or spherical harmonic decompositions.

A well-known property of the analysis process is its ability to filter noise from observations. By defining the filter operator \( \bar{F} \equiv H \bar{K}, \) for the case of a linear observation operator, we can write the analysis in observation space as
\[ H x_a = (I - H \bar{K}) H x_b + H K y = (I - \bar{F}) H x_b + \bar{F} y. \] (4)

Equation (4) shows that, in the linear case, the analysis in observation space is obtained by operating with \( \bar{F} \) on the observation vector and operating with \( I - \bar{F} \) on the background state (in observation space). In the context of remote sensing, the operator \( \bar{F} \) is related to the averaging kernel \( A = \bar{K} H \) by \( HA = \bar{F} H. \) The operators \( \bar{F} \) and \( I - \bar{F} \) can be expressed in terms of the simultaneous eigendecomposition of \( \bar{H} \bar{B} \bar{H}^T \) and \( \bar{R} \) via Equations (3a) and (3b) as
\[ \bar{F} = \bar{H} \bar{B} \bar{H}^T \bar{D}^{-1} = S^{-T} (I + \Phi)^{-1} S^T, \] (5a)
\[ I - \bar{F} = \bar{R} \bar{D}^{-1} = S^{-T} \Phi (I + \Phi)^{-1} S^T. \] (5b)

As noted in Hollingsworth (1986), it is apparent that \( \bar{F} \) is a filter that dampens modes prominent in \( \bar{R} \) as compared to \( \bar{H} \bar{B} \bar{H}^T \), while \( I - \bar{F} \) dampens modes prominent in \( \bar{H} \bar{B} \bar{H}^T \) as compared to \( \bar{R} \); modes where the simultaneous eigenvalues of \( \bar{H} \bar{B} \bar{H}^T \) are significantly larger than those of \( \bar{R} \) correspond to elements in \( \Phi \) that are much smaller than unity, so the corresponding elements in \( (I + \Phi)^{-1} \) and \( \Phi (I + \Phi)^{-1} \) will be close to unity and zero, respectively, and
vice versa for modes where the simultaneous eigenvalues of $\hat{R}$ are significantly larger than $\hat{H}\hat{B}H^T$ (Figure 1 bottom plots give an illustration).

Error statistics tuning algorithms that make use of innovation statistics (either directly or indirectly via assimilation) must in some way differentiate which parts of the innovation statistics can be attributed to the background and which to the observations. When explicitly demonstrating how the DI01 and D05 algorithms make this separation, it will prove useful to introduce the following notation related to the filter $\tilde{F}$ used to generate the analysis. For a matrix $A$, we define the matrices $A_B$ and $A_R$ as the matrix $A$ filtered by $F$ and $I - \tilde{F}$, respectively, so that

$$A_B \equiv \tilde{F}AP^T,$$
$$A_R \equiv (I - \tilde{F})A(I - \tilde{F})^T.$$

The subscripts "$B$" and "$R$" on these quantities are to signify that $A_B$ only retains the modes in $A$ that are prominent in $\hat{H}\hat{B}H^T$ and $A_R$ only retains the modes in $A$ that are prominent in $\hat{R}$ (tildes on the subscripts are implied but not explicitly written). With this, we define the filtered observed innovation covariances $D_B$ and $D_R$ as

$$D_B \equiv FDF^T,$$
$$D_R \equiv (I - \tilde{F})D(I - \tilde{F})^T$$

and their modelled analogues $\tilde{D}_B$ and $\tilde{D}_R$ as

$$\tilde{D}_B \equiv \tilde{F}D\tilde{F}^T = \hat{H}\hat{B}H^TD^{-1}\hat{H}\hat{B}H^T,$$
$$\tilde{D}_R \equiv (I - \tilde{F})\tilde{D}(I - \tilde{F})^T = \hat{R}\tilde{D}^{-1}\hat{R}.$$

Filtering the innovation covariance using $\tilde{F}$ and $I - \tilde{F}$ offers a means to isolate the contribution of the background from the observations (Figure 6 below shows an example of the background-filtered innovation covariance).

For the BLUE analysis, the analysis-minus-background difference in observation space, which approximately equals the analysis increment in

**FIGURE 1** Modelled background- and observation-error covariances as a function of wavenumber. In each panel, the top plots show the covariances and the bottom plots show the filter functions. Covariances are homogeneous and defined on a 1D periodic domain with 1001 equally spaced grid points. Background correlations are modelled using a Gaussian with correlation length $L_b$ and observation errors are uncorrelated in space. The units of the covariances are arbitrary.
observations space, is given by $H(x_a) - H(x_b) \approx H\delta x_a = \delta d$. Similarly, the observation-minus-analysis difference, which we can also refer to as the *a posteriori* observation residual, is given by $y - H(x_a) \approx d - H\delta x_a = (1 - \tilde{F})d$. The filtered observed innovation covariances $D_B$ and $D_R$ can then be expressed in terms of these differences as

$$D_B = Fdd^T T \approx \{H(x_a) - H(x_b)\} \{H(x_a) - H(x_b)\}^T,$$

(9a)

$$D_R = (I - \tilde{F})dd^T (I - \tilde{F})^T \approx \{y - H(x_a)\} \{y - H(x_a)\}^T.$$

(9b)

As such, we may refer to $D_B$ and $D_R$ as the analysis increment covariance and the *a posteriori* observation residual covariance, respectively.

### 4 THE FROBENIUS INNER PRODUCT

Another important concept, which we will use extensively in this paper, is the Frobenius inner product. An inner product space associates a scalar quantity to a pair of vectors defined in a vector space. The most common inner product is the ubiquitous dot product. This concept can be extended to matrices by use of the vectorization operator $vec(\cdot)$, which converts a matrix into a column vector by stacking the columns of the matrix on top of one another. The Frobenius inner product between the real-valued matrices $A$ and $B$, denoted by $\langle A, B \rangle$, is defined as the dot product between the vectorizations of the matrices, so that

$$\langle A, B \rangle = \sum_{ij} A_{ij}B_{ij} = \text{vec}(A)^T \text{vec}(B) = \text{Tr}[A^T B].$$

(10)

In this way, the Frobenius inner product uses the trace operator to map two matrices into a scalar. The Frobenius product of a matrix with itself defines the square of the Frobenius norm, which for a matrix $A$ is denoted by $\|A\| = \langle A, A \rangle^{1/2}$. As a norm, the Frobenius norm is always greater or equal to zero.

As with other inner products, we can extend the definition of the Frobenius inner product to include a weighting. Although we may weight the inner product by any positive definite matrix, for this work we restrict ourselves to weightings that can be written in the form $W_1 \otimes W_2$, where $W_1$ and $W_2$ are (symmetric) positive definite matrices and $\otimes$ denotes the Kronecker product. The weighted Frobenius inner product is then given by

$$\langle A, B \rangle_{W_1 \otimes W_2} = \text{vec}(A)^T (W_1 \otimes W_2) \text{vec}(B) = \text{Tr}[A^T W_2 B W_1] = \langle W_2 A, B W_1 \rangle,$$

(11)

where in the second equality we have made use of Equation (B9) from Appendix B. We direct the reader to Appendix B for a review of the Kronecker product, the vectorization operator, and some related identities. In the last equality of Equation (11), we see that, by choosing a weighting of the form $W_1 \otimes W_2$, the weighted Frobenius inner product can be written in terms of the unweighted product by weighting each factor in the product separately. The weighted Frobenius norm of matrix $A$ weighed by $W_1 \otimes W_2$ is given by $\|A\|_{W_1 \otimes W_2} = \langle A, A \rangle_{W_1 \otimes W_2}^{1/2}$.

As with the Euclidean inner product, we can decompose the (weighted) Frobenius inner product into norms and an angle defined with respect to the Frobenius inner product. The angle $\theta_{A,B}$ between matrices $A$ and $B$ with respect to the weighted Frobenius inner product is defined by

$$\cos(\theta_{A,B}) = \frac{\langle A, B \rangle_{W_1 \otimes W_2}}{\|A\|_{W_1 \otimes W_2} \cdot \|B\|_{W_1 \otimes W_2}}.$$

(12)

In this paper both inner products and expectation values will be discussed. As in some references the angled brackets $\langle \cdot \rangle$ are used to denote the expectation value, we use the notation $E \{\cdot\}$ instead to denote the expectation value and caution the reader not to conflate these two operations.

We now proceed to discuss three error estimation methods in Sections 5, 6, and 7, and reformulate them in terms of minimization of cost functions.

### 5 THE DESROZIERS AND IVANOV (2001) METHOD

The first error covariance tuning method we will examine is the tuning method of Desroziers and Ivanov (2001) (DI01). While the DI01 method can be difficult to implement in practice due to the need to evaluate the trace of a large non-diagonal matrix (Chapnik et al., 2006 for possible operational implementations), the insights gained by examining the DI01 method will carry over to other methods examined in this paper.

#### 5.1 Description of the algorithm

In this section, we give a brief review of the DI01 algorithm. The DI01 algorithm is rooted in a comparison of the expectation values of the background and observation cost functions evaluated at the analysis to their actual values as calculated within an assimilation. The expectation values of $J_b$ and $J_o$ evaluated at the BLUE analysis can
be found via Equation (B11) of Appendix B as

\[
E[J_b(x_o)] = \frac{1}{2} \text{Tr} \left[ H B H^T D^{-1} E[D] \right]
\]

\[
E[J_o(x_o)] = \frac{1}{2} \text{Tr} \left[ R \tilde{D}^{-1} E[D] \tilde{D}^{-1} \right]
\]

(13a)

(13b)

As noted in Talagrand (1999); Talagrand (2010), these expectation values simplify considerably if the modelled innovation covariance \( D \) equals the expectation value of \( D \), a case known as innovation covariance consistency (Ménard, 2016), which are shown in Equations (13a) and (13b) to the right of the arrows. Note that the rightmost expressions in Equations (13a) and (13b) depend on the modelled quantity \( H K \) as opposed to its true value.

In practice, diagnostic relations are found by replacing the expectation values \( E[J_b(x_o)] \) and \( E[J_o(x_o)] \) in Equations (13a) and (13b) with their actual observed quantities. These diagnostic relations are only a consistency check on the error statistics, but not yet an algorithm by which the error statistics can be adjusted. The DI01 method establishes an iterative tuning method by assigning an iteration number to each modelled value (i.e., each term with a tilde), making the choice of iteration values

\[
\text{Tr} \left[ H B H^T \tilde{D}_i^{-1} \tilde{D}_i \right] = \text{Tr} \left[ H B_{i+1} H^T \tilde{D}_i \right],
\]

\[
\text{Tr} \left[ \tilde{R} \tilde{D}_i^{-1} \tilde{D}_i \right] = \text{Tr} \left[ \tilde{R}_{i+1} \tilde{D}_i^{-1} \right],
\]

(14a)

(14b)

where the subscript \( i \) denotes values for the \( i \)th iteration of the algorithm. The error covariances are then parametrized through two scaling parameters \( s_b \) and \( s_o \), in which the error covariances \( B_{i+1} \) and \( \tilde{R}_{i+1} \) at iteration \( i+1 \) are computed from their values at the previous iteration by \( B_{i+1} = (s_b)_{i+1} B_i \) and \( \tilde{R}_{i+1} = (s_o)_{i+1} \tilde{R}_i \). The DI01 scaling coefficients at iteration \( i+1 \), \( s_b^{DI01} \) and \( s_o^{DI01} \), can then be solved as

\[
(s_b^{DI01})_{i+1} = \frac{\text{Tr} \left[ H B H^T \tilde{D}_i^{-1} \tilde{D}_i \right]}{\text{Tr} \left[ H B H^T \tilde{D}_i \right]},
\]

\[
(s_o^{DI01})_{i+1} = \frac{\text{Tr} \left[ \tilde{R} \tilde{D}_i^{-1} \tilde{D}_i \right]}{\text{Tr} \left[ \tilde{R} \tilde{D}_i \right]},
\]

(15a)

(15b)

and the initial scaling values of \( (s_b^{DI01})_{i=0} = 1 \) and \( (s_o^{DI01})_{i=0} = 1 \) are chosen. The DI01 algorithm was designed for use within a variational assimilation system, where the numerators of the scaling parameters in Equations (15a) and (15b) would have already been calculated in the cost function computation required for a variational system. Additional computations must be done for the denominators which, for example, can be calculated by means of a Monte Carlo estimate of the trace (Chapnik et al., 2006). For the rest of the paper, we drop the iteration number subscript to simplify notation.

The DI01 algorithm was only formulated to compute variance scaling factors, either global scalings or for various data subsets defined by projection operators (Chapnik et al., 2004). Although Equations (14a) and (14b) can be used to accommodate other types of parametrization (for instance for correlation lengths), finding an explicit analytical solution for these parametrizations may be difficult.

## 5.2 Associated cost functions

We now aim to associate the variance scaling parameters of the DI01 algorithm to cost functions that when minimized yield the iterative relations in Equations (15a) and (15b). To this end, using the filtered innovation covariances introduced in the previous section, we rewrite the expressions for \( s_b^{DI01} \) and \( s_o^{DI01} \) as

\[
s_b^{DI01} = \frac{\text{Tr} \left[ \tilde{D}_b \tilde{D}_b^{-1} \tilde{D}_b (H B H^T)^{-1} \right]}{\text{Tr} \left[ \tilde{D}_b \tilde{D}_b^{-1} \tilde{D}_b (H B H^T)^{-1} \right] \text{vec}(\tilde{D}_b)^T (\tilde{D}_b \otimes (H B H^T)^{-1})^{-1} \text{vec}(\tilde{D}_b)},
\]

(16a)

\[
s_o^{DI01} = \frac{\text{Tr} \left[ \tilde{D}_o \tilde{D}_o^{-1} \tilde{D}_o (H B H^T)^{-1} \right]}{\text{Tr} \left[ \tilde{D}_o \tilde{D}_o^{-1} \tilde{D}_o (H B H^T)^{-1} \right] \text{vec}(\tilde{D}_o)^T (\tilde{D}_o \otimes \tilde{R})^{-1} \text{vec}(\tilde{D}_o)},
\]

(16b)

where we have used Equations (B3) and (B9) in deriving the second lines of these equations. From the second lines of Equations (16a) and (16b), it is apparent that \( s_b^{DI01} \) and \( s_o^{DI01} \) have the form of linear least-squares coefficients (as may be seen in any general reference on linear regression analysis, e.g., chapter 3.10 of Seber and Lee 2003). These coefficients minimize the cost functions \( J_B^{DI01} \) and \( J_R^{DI01} \) which are defined by

\[
J_B^{DI01} = \frac{1}{2} \text{vec} \left( s_b \tilde{D}_B - D_B \right)^T (\tilde{D}_B \otimes (H B H^T)^{-1})^{-1} \text{vec} \left( s_b \tilde{D}_B - D_B \right)
\]

\[
\times \text{vec} (s_b \tilde{D}_B - D_B)
\]

\[
= \frac{1}{2} \text{Tr} \left[ (s_b \tilde{D}_B - D_B) \tilde{D}_B^{-1} (s_b \tilde{D}_B - D_B) (H B H^T)^{-1} \right]
\]
where we have used Equations (11) to write the cost functions in terms of the trace operator and the Frobenius inner product in addition to the vectorization operator. Appendix C gives details on the minimization of these cost functions. As a notational note, the cost functions $J_{B}^{D01}$ and $J_{R}^{D01}$ are written as a function of $s_{B}$ and $s_{o}$, respectively, while $s_{b}^{D01}$ and $s_{o}^{D01}$ denote the values of $s_{B}$ and $s_{o}$ that minimize $J_{B}^{D01}$ and $J_{R}^{D01}$, respectively. The cost functions $J_{B}^{D01}$ and $J_{R}^{D01}$ have the familiar form of cost functions that measure the squared norm of the data vectors $\text{vec}(s_{B} \mathbf{D}_{B} - \mathbf{D}_{B})$ and $\text{vec}(s_{o} \mathbf{D}_{R} - \mathbf{D}_{R})$, weighted by $(\mathbf{D}_{B} \otimes (\mathbf{H} \mathbf{B}^{T}))^{-1}$ and $(\mathbf{D}_{R} \otimes \mathbf{R})^{-1}$, respectively. The variance scaling parameter $s_{B}^{D01}$ is the linear least-squares coefficient for fitting the prescribed background-filtered innovation covariance $\mathbf{D}_{B}$ to its observed value $\mathbf{D}_{B}$, and similarly $s_{o}^{D01}$ is the linear least-squares coefficient for fitting the observation-filtered quantities. It is now clear that the DI01 method implicitly uses the filter $\hat{\mathbf{F}}$ that produces the analysis in observation space (through Equation (4)) to separate the background from the observational contributions to the innovation covariance.

If an analysis has already been produced using a variational assimilation system, by using the same filter that produces the analysis, the DI01 algorithm does not need to explicitly compute the contributions of $\mathbf{D}_{B}$ and $\mathbf{D}_{R}$ to the tuning parameters as this would have already been computed during the evaluation of the cost functions in the variational system. Although one could envision using different filter operators in $J_{B}^{D01}$ and $J_{R}^{D01}$, using that same filter as used to produce the analysis significantly decreases the computational cost of the algorithm. As the DI01 algorithm is reliant on the prescribed filter operator being able to adequately separate the background from the observational contributions to the innovation covariance, the algorithm will only be successful if this filtering is done to a sufficiently accurate degree, which will be quantified in Section 8.

In this section, we have seen that the DI01 method is equivalent to the minimization of the cost functions $J_{B}^{D01}$ and $J_{R}^{D01}$, defined in Equations (17a) and (17b). By casting the DI01 method in terms of the minimization of particular cost functions, we can directly compare the DI01 method to other algorithms by comparing cost functions, provided we know the cost functions associated with the other method. Accordingly, we will derive the cost functions associated with the D05 and HL methods in the following two sections in order to make these comparisons.

### 6 THE DESROZIERS ET AL. (2005) METHOD

Although the DI01 algorithm may be used to tune error variances, the evaluation of traces required is often computationally demanding in operation settings. In addition, the DI01 method as originally devised only allows for the tuning of error variances, while other covariance parameters, such as correlation lengths, are not updated. An alternative method for tuning error statistics was devised by Desroziers et al. (2005) which is typically less computationally demanding and can be used to tune correlation parameters. In this section, we examine the D05 method and associate quadratic cost functions with it, which allows for a direct comparison between the DI01 and D05 algorithms. As we wish to compare the DI01 and D05 methods, we concentrate on scaling parameters for the error variances, but note that the methods presented in this section can be generalized for other tuning parameters.

#### 6.1 Description of the algorithm

D05 notes that one can form various consistency checks for the error covariances by taking the expectation value of various outer products of differences in observation space, which include the relations

$$E \left[ (\mathbf{H}(\mathbf{x}_{a}) - \mathbf{H}(\mathbf{x}_{b}))^{T} \mathbf{D} \right] = \mathbf{H}^{T} \mathbf{D}^{-1} \mathbf{E}[\mathbf{D}]^{\mathbf{D}} \mathbf{H}^{T},$$  

(18a)

$$E \left[ (\mathbf{y} - \mathbf{H}(\mathbf{x}_{b}))^{T} \mathbf{D} \right] = \mathbf{R} \mathbf{D}^{-1} \mathbf{E}[\mathbf{D}]^{\mathbf{D}} \mathbf{R}.$$  

(18b)

If the modelled innovation covariance $\mathbf{D}$ is equal to the expectation value of the observed innovation covariance $\mathbf{D}$, then the expectation values in Equations (18a) and (18b) simply become equal to $\mathbf{H}^{T} \mathbf{D}^{T}$ and $\mathbf{R}$. Based on these relations, D05 proposed that the iterative relations

$$\mathbf{H}_{1}^{T} \mathbf{H}^{T} = \mathbf{H}^{T} \mathbf{D}^{-1} \mathbf{D},$$  

(19a)

$$\mathbf{R}_{t+1} = \mathbf{R} \mathbf{D}^{-1} \mathbf{D},$$  

(19b)

be used as a tuning algorithm. Ménard (2016) showed that, when the full matrices of both $\mathbf{H}^{T} \mathbf{D}^{T}$ and $\mathbf{R}$ are updated with this algorithm, innovation consistency is reached after a single iteration.

The D05 algorithm as given in Equations (19a) and (19b) describes the tuning of the full covariance matrices.
where we have allowed for the (as of yet unspecified) weightings $\mathbf{W}_{b_1} \otimes \mathbf{W}_{b_2}$ and $\mathbf{W}_{o_1} \otimes \mathbf{W}_{o_2}$. By selecting the covariance parameters in this manner, the parameters are the best fit parameters corresponding to the cost functions $J_B$ and $J_R$. We note that, contrary to the DI01 cost functions in Equations (17a) and (17b), as the cost functions in Equations (21a) and (21b) do not specify a particular parametrization, they can be used to estimate error correlations as well.

Now restricting ourselves to scalings of the variances, these cost functions simplify to

$$J_B = \frac{1}{2} \| \mathbf{H}\mathbf{B}^T \mathbf{D}^{-1} \mathbf{D} - \mathbf{H}\mathbf{B}^T \mathbf{D}^{-1} \mathbf{D} \|^2_{\mathbf{W}_{b_1} \otimes \mathbf{W}_{b_2}}$$

$$J_R = \frac{1}{2} \| \mathbf{I} - \mathbf{F}^T \mathbf{D}^{-1} \mathbf{D} \|^2_{\mathbf{W}_{o_1} \otimes \mathbf{W}_{o_2}}$$

where Equations (11), (B4) and (B6) have been used in the third lines to move the filter operators from inside the norm to the weighting. At this point, as with the DI01 method, we have dropped the iteration index to simplify notation.

These cost functions are minimized by the scaling parameters

$$\hat{s}_b = \frac{\text{Tr} \left[ \mathbf{W}_{b_1} \mathbf{H}\mathbf{B}^T \mathbf{W}_{b_2} \mathbf{H}\mathbf{B}^T \mathbf{D}^{-1} \mathbf{D} \right]}{\text{Tr} \left[ \mathbf{W}_{b_1} \mathbf{H}\mathbf{B}^T \mathbf{W}_{b_2} \mathbf{H}\mathbf{B}^T \right]}$$

$$\hat{s}_o = \frac{\text{Tr} \left[ \mathbf{W}_{o_1} \mathbf{R} \mathbf{W}_{o_2} \mathbf{R} \mathbf{D}^{-1} \mathbf{D} \right]}{\text{Tr} \left[ \mathbf{W}_{o_1} \mathbf{R} \mathbf{W}_{o_2} \mathbf{R} \right]}$$

Note that these expressions have the scaling parameter estimates written with the accent ^ instead of having the superscript D05, as this superscript will be reserved for the particular choice of weightings chosen in Equations (20a) and (20b). Comparing Equations (23a) and (23b) to Equations (20a) and (20b), it is apparent that the scaling coefficients in the standard D05 algorithm can be recovered by choosing the weightings as $\mathbf{W}_{b_1} = (\mathbf{H}\mathbf{B}^T)^{-1}$, $\mathbf{W}_{b_2} = \mathbf{I}$, $\mathbf{W}_{o_1} = \mathbf{R}^{-1}$, and $\mathbf{W}_{o_2} = \mathbf{I}$. Note that this choice of weighting is not unique. For instance, the weightings $\mathbf{W}_{b_1} = \mathbf{W}_{b_2} = (\mathbf{H}\mathbf{B}^T)^{-1/2}$ and $\mathbf{W}_{o_1} = \mathbf{W}_{o_2} = \mathbf{R}^{-1/2}$ yield the D05 relations as well, where $(\mathbf{H}\mathbf{B}^T)^{1/2}$ and $\mathbf{R}^{1/2}$ are
defined to be the symmetric square roots of each covariance matrix. With this choice of weightings, the cost functions become

\[
\begin{align*}
J^\text{D05}_R &= \frac{1}{2} \left\| \hat{s}_B \mathbf{D}_B - \mathbf{D}_R \right\|^2_{(R_R \otimes I)^{-1}} \quad (24a) \\
J^\text{D05}_R &= \frac{1}{2} \left\| \hat{s}_R \mathbf{D}_R - \mathbf{D}_R \right\|^2_{(R_R \otimes I)^{-1}}, \quad (24b)
\end{align*}
\]

where we have defined \( \hat{\mathbf{B}}_B \equiv \mathbf{FHF}^T \mathbf{F}^T \) and \( \hat{\mathbf{R}}_R \equiv (\mathbf{I} - \hat{\mathbf{F}})(\mathbf{I} - \hat{\mathbf{F}})^T \) via Equations (6a) and (6b), respectively (\( \hat{\mathbf{B}}_B \) will be written without \( \mathbf{H} \) or \( \mathbf{H}^T \) for ease of notation). Comparing \( J^\text{D05}_B \) and \( J^\text{D05}_R \) in Equations (24a) and (24b) to \( J^\text{DI01}_B \) and \( J^\text{DI01}_R \) in Equations (17a) and (17b) for the DI01 method, we see that, for the scaling of the variances, the two methods differ only in their weighting of the cost functions. Therefore, although the DI01 and D05 algorithms have important differences in their numerics and ease of implementation, conceptually both methods are solutions to essentially the same problem (minimization of the cost functions associated with the filtered innovation covariances) with only minor differences (the weighting factors in the cost functions). In other words, both DI01 and D05 methods implicitly fit a modelled covariance to an observed covariance, but give a different weighting to each term in the covariance matrix. The differences in assigned weightings between the two methods is further explored in Section 11. As we have parametrized the modelled covariances with scaling factors, the implicit fitting procedure is simply a linear least-squares fit. Also note that, with this choice of weighting, \( J^\text{D05}_B \) and \( J^\text{D05}_R \) have the same units as \( \mathbf{HH}^T \) and \( \hat{\mathbf{R}} \), unlike \( J^\text{DI01}_B \) and \( J^\text{DI01}_R \) which are unitless (although the identity matrices in the weightings can be made to include units so that \( J^\text{D05}_B \) and \( J^\text{D05}_R \) are unitless as well).

The weightings chosen in Equations (24a) and (24b) are only one possible choice and, by choosing different weightings, one can produce different iterative schemes for the variance scaling coefficients. Each choice of weightings will impact the performance of the iterative tuning scheme. For instance, with the choice of \( \mathbf{W}_{b1} = \mathbf{W}_{b2} = (\mathbf{HH}^T)^{-1} \) and \( \mathbf{W}_{o1} = \mathbf{W}_{o2} = \mathbf{R}^{-1} \), the scheme gives the relations \( \hat{s}_B = \text{Tr}[\mathbf{DD}^T]/N_{\text{obs}} \) and \( \hat{s}_R = \text{Tr}[\mathbf{DD}^T]/N_{\text{obs}} \), where \( N_{\text{obs}} \) is the number of observations. In this case, the ratio between the background- and observation-error variances would not change between iterations, thereby not affecting the analyses made using the updated error covariances. With this choice of weightings, this method simplifies to an algorithm that satisfies the \( \chi^2 \) diagnostic without changing the relative scaling between the background- and observation-error covariances. If instead the weightings are chosen as \( \mathbf{W}_{b1} = \mathbf{D}_B^{-1}, \mathbf{W}_{b2} = (\mathbf{HH}^T)^{-1}, \mathbf{W}_{o1} = \mathbf{D}_R^{-1}, \) and \( \mathbf{W}_{o2} = \mathbf{R}^{-1} \), then the DI01 method is recovered. As such, for the case of variance scaling, both the DI01 and D05 methods belong to a larger class of statistics tuning methods defined by Equations (22a) and (22b).

As with the DI01 method, we have seen that the tuning of variances using the D05 method is equivalent to the minimization of quadratic cost functions. Additionally, both the DI01 and D05 methods use the analysis filter to attribute portions of the innovation covariance to the background and observations. A comparison of their associated cost functions shows that, fundamentally, the two algorithms differ only by the weightings used in each cost function. The DI01 and D05 methods both belong to the class of methods defined by Equations (22a) and (22b), in which cost functions measuring the mismatch between parts of the modelled and observed innovation covariances are minimized, using the analysis filter for attribution to either the background or observations.

7 | THE HOLLINGWORTH–LÖNNBERG METHOD

The method of Hollingsworth and Lönnberg (1986) provides a method for diagnosing error covariances through the use of innovation statistics. But unlike the DI01 and D05 algorithms, the HL method does not use the filter that is used to produce the analysis to differentiate between the observation and background contributions to the innovation covariance. Instead, the observation errors are assumed to be spatially uncorrelated, so all spatial correlations in the innovation covariance are attributed to the background. As the analysis filter is not used, the HL method is an \textit{a priori} scheme.

7.1 | Description of the algorithm

Unlike the DI01 and D05 methods, the HL method determines the background and observation parameters sequentially. First the background-error covariance parameters are determined by fitting the innovation covariance matrix without the diagonal terms so as to utilize only the parts of the innovation covariance affected solely by the background. The modelled best-fit parameters are then used to extrapolate the background-error covariance to zero spatial separation, so that the background contribution to the innovation variance can be removed and an estimate of the observation-error variance can be made.

The HL method typically parametrizes both the variances and correlations of the background covariance. However, to compare with the DI01 and D05 methods,
we restrict ourselves to a single global background-error variance scaling parameter $s_b^{\text{HL}}$. The first step in the HL method fits the off-diagonal elements of $\mathbf{HBH}^T$, which are the same as the off-diagonal elements of $\tilde{\mathbf{D}}$ since the observation errors are assumed to be uncorrelated. As in Ménard et al. (2016), this fitting procedure can be done via the minimization of the cost function $J_B^{\text{HL}}$, defined as

$$J_B^{\text{HL}} = \frac{1}{2} \sum_{ij \neq i} \left( s_b \mathbf{C} - \tilde{\Sigma}_b^{-1} \mathbf{D} \tilde{\Sigma}_b^{-1} \right)_{ij}^2,$$  

(25)

where $i, j$ are matrix indices, $\tilde{\Sigma}_b$ is a diagonal weighting matrix, and $\mathbf{C}$ is the modelled background-error correlation matrix. We take $\tilde{\Sigma}_b$ as the diagonal matrix of square roots of the modelled background variances, so that the modelled background covariance can be expressed as $\mathbf{HBH}^T = \tilde{\Sigma}_b \mathbf{C} \tilde{\Sigma}_b$. An estimate of the background-error variance can then be made using this fit and can then be subtracted from the observed innovation-error variance to estimate the observation-error variance, which may be performed as

$$s_b^{\text{HL}} = \frac{1}{N_{\text{obs}}} \sum_i \left( \mathbf{D} - s_b^{\text{HL}} \mathbf{HBH}^T \right)_{ii},$$  

(26)

where $s_b^{\text{HL}}$ is the observation-error variance scaling parameter for the HL method and $s_b^{\text{HL}}$ is the previously determined best-fit parameter for $s_b$.

## 7.2 Associated cost functions

Although the previous section introduced the HL background cost function $J_B^{\text{HL}}$, its form in Equation (25) is difficult to compare with the cost functions for the DI01 and D05 methods, and we have yet to introduce a cost function for the observation-error covariance. In this section, we complete the formulation of the HL method in terms of the minimization of cost functions that measure the mismatch between the modelled and observed background- and observation-error covariances.

By introducing the constant $N_{\text{obs}} \times N_{\text{obs}}$ matrix $\mathbf{V}$ that has all elements equal to one, except the diagonal elements that are equal to zero, the cost function $J_B^{\text{HL}}$ in Equation (25) can be written as

$$J_B^{\text{HL}} = \frac{1}{2} \sum_{ij \neq i} \left( \tilde{\Sigma}_b^{-1} (s_b \mathbf{D} - \mathbf{D}) \tilde{\Sigma}_b^{-1} \right)_{ij}^2$$

$$= \frac{1}{2} \sum_{ij} \left( \tilde{\Sigma}_b^{-1} (s_b \mathbf{V} \mathbf{D} - \mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-1} \right)_{ij}^2$$

$$= \frac{1}{2} \text{Tr} \left[ (s_b \mathbf{V} \mathbf{D} - \mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} (s_b \mathbf{V} \mathbf{D} - \mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} \right]$$

$$= \frac{1}{2} \left\| s_b \mathbf{V} \mathbf{D} - \mathbf{V} \mathbf{D} \right\|^2 \left( \tilde{\Sigma}_b \tilde{\Sigma}_b^{-1} \right)^{-1},$$  

(27)

where $\circ$ denotes the Hadamard or element-wise product (Equation (B10)) and Equation (11) has been used in the last line. Introducing the matrix $\mathbf{V}$ with the Hadamard product allows us to sum over all matrix indices and consequently $J_B^{\text{HL}}$ can be expressed in terms of a trace, allowing for a direct comparison with the DI01 and D05 methods. The most notable difference between $J_B^{\text{HL}}$ and its analogues for the DI01 and D05 methods, found in Equations (17a) and (24a), respectively, is the terms involving the filtered innovation covariances, $\mathbf{D}_b$ and $\tilde{\mathbf{D}}_b$, are replaced by $\mathbf{V} \mathbf{D}$ and $\mathbf{V} \tilde{\mathbf{D}}$ in $J_B^{\text{HL}}$, respectively. As such, the element-wise multiplication between the matrix $\mathbf{V}$ and the innovation covariance is analogous to the filtering that is done using the operator $\mathbf{F}$ in the DI01 and D05 methods. The weighting of $\left( \tilde{\Sigma}_b \circ \tilde{\Sigma}_b \right)^{-1}$ chosen for $J_B^{\text{HL}}$ also differs from the weightings used in $J_B^{\text{DI01}}$ or $J_B^{\text{D05}}$ but we note that $J_B^{\text{HL}}$ could be defined with a more complicated weighting than that used here. The value of $s_b$ that minimizes $J_B^{\text{HL}}$ is given by

$$s_b^{\text{HL}} = \frac{\text{Tr} \left[ (\mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} (\mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} \right]}{\text{Tr} \left[ (\mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} (\mathbf{V} \mathbf{D}) \tilde{\Sigma}_b^{-2} \right]} = \frac{\left\| \mathbf{V} \mathbf{D} \right\|^2 \left( \tilde{\Sigma}_b \tilde{\Sigma}_b^{-1} \right)^{-1}}{\left\| \mathbf{V} \mathbf{D} \right\|^2 \left( \tilde{\Sigma}_b \tilde{\Sigma}_b^{-1} \right)^{-1}}.$$  

(28)

After the background-error covariance parameters have been determined, a fit for the observation-error variance parameters can be made using the diagonal elements of the innovation covariance matrix. We define the cost function $J_B^{\text{HL}}$ as

$$J_B^{\text{HL}} = \frac{1}{2} \sum_i \left( s_b^{-1} (s_b \mathbf{HBH}^T + s_b \mathbf{R} - \mathbf{D}) \tilde{\Sigma}_o^{-1} \right)_{ii}^2$$

$$= \frac{1}{2} \sum_{ij} \left( \tilde{\Sigma}_o^{-1} (\mathbf{I} \circ (s_b^{\text{HL}} \mathbf{HBH}^T + s_b \mathbf{R}) - \mathbf{I} \circ \mathbf{D}) \tilde{\Sigma}_o^{-1} \right)_{ij}^2$$

$$= \frac{1}{2} \text{Tr} \left[ (\mathbf{I} \circ (s_b^{\text{HL}} \mathbf{HBH}^T + s_b \mathbf{R}) - \mathbf{I} \circ \mathbf{D}) \tilde{\Sigma}_o^{-2} \times (\mathbf{I} \circ (s_b^{\text{HL}} \mathbf{HBH}^T + s_b \mathbf{R}) - \mathbf{I} \circ \mathbf{D}) \tilde{\Sigma}_o^{-2} \right]$$

$$= \frac{1}{2} \left\| \mathbf{I} \circ (s_b^{\text{HL}} \mathbf{HBH}^T + s_b \mathbf{R}) - \mathbf{I} \circ \mathbf{D} \right\|_{\left( \tilde{\Sigma}_o \circ \tilde{\Sigma}_o^{-1} \right)^{-1}},$$  

(29)

where $\tilde{\Sigma}_o$ is a diagonal weighting matrix. As the fits for the background- and observation-error covariance parameters are done sequentially, the best-fit background parameter $s_b^{\text{HL}}$ appears in the cost function $J_B^{\text{HL}}$. In Equation (29), the Hadamard product is again used to extend the sum over all matrix indices, this time by an element-wise multiplication with the identity matrix (of dimension $N_{\text{obs}} \times$...
As with $J_R^{HL}$, by comparing $J_R^{D05}$ with its analogues for the DI01 and D05 methods, found in Equations (17b) and (24b), respectively, we see that the Hadamard product with the identity matrix in $J_R^{HL}$ is analogous to the filtering done with the operator $I - \bar{F}$ in the DI01 and D05 methods. The cost function $J_R^{HL}$ is minimized by the scaling coefficient

$$
J_{o}^{HL} = \frac{\text{Tr} \left[ (I_o \bar{R}) \bar{\Sigma}_o^{-2} (I_o \left( D - s_b^{HL} H BH^T \right) ) \bar{\Sigma}_o^{-2} \right]}{\text{Tr} \left[ (I_o \bar{R}) \bar{\Sigma}_o^{-2} (I_o \bar{R}) \bar{\Sigma}_o^{-2} \right]}
= \frac{\langle I_o \bar{R} (I_o (D - s_b^{HL} H BH^T)) \bar{\Sigma}_o^{-1} \rangle}{\| I_o \bar{R} \|^2 (\bar{\Sigma}_o^{-1} \otimes \bar{\Sigma}_o^{-1})^{-1}}.
$$

(30)

If the weighting $\bar{\Sigma}_o$ is chosen to be the diagonal matrix of the square roots of the modelled observation-error variances, as we do for the remainder of the paper, then Equation (30) simplifies to Equation (26).

By using the Hadamard product with the matrix $V$ or $I$, we have seen that we can construct cost functions for the HL method that can be directly compared to the cost functions associated with the DI01 method (Equations (17a) and (17b)) or with the D05 method (Equations (24a) and (24b)). By comparing their associated cost functions, all three methods are associated with cost functions quadratic in the variance scaling parameters and all three methods have different weightings in their associated cost functions. While the DI01 and D05 methods use the analysis filter to attribute portions of the innovation covariance to either the background or observations, the HL method attributes all correlations to the background.

## 8 | Geometric Interpretations of Covariances

We have seen that both the DI01 and D05 a posteriori algorithms rely on the filter used to produce the analysis to separate the contributions of the background and observations to the innovation covariance and are a posteriori algorithms precisely due to the use of this filter. The ability of the filter to accurately perform this separation is crucial for successful applications of the DI01 and D05 methods. This will be dependent on how different the background- and observation-error covariance spectra are from one another. This section introduces measures to help quantify the degree of difference between the background- and observation-error covariances, which will be done using the Frobenius inner product. The use of the Frobenius inner product will also allow us to interpret the covariance matrices geometrically.

### 8.1 Inner Products Between Covariances

We begin this section by returning to Sections 5.1 and 6.1 and rewriting the expressions for the variance scaling parameters in terms of the Frobenius inner product. For the DI01 method, Equations (15a) and (15b) take the form

$$
x_b^{DI01} = \frac{\langle HH^T, D \rangle_{(D \otimes D)^{-1}}}{\langle HH^T, \bar{D} \rangle_{(D \otimes D)^{-1}}},
$$

(31a)

$$
x_o^{DI01} = \frac{\langle \bar{R}, D \rangle_{(D \otimes D)^{-1}}}{\langle \bar{R}, \bar{D} \rangle_{(D \otimes D)^{-1}}},
$$

(31b)

while Equations (20a) and (20b) for the D05 method are written

$$
x_b^{D05} = \frac{\langle HH^T, D \rangle_{(D \otimes D)^{-1}}}{\langle HH^T, \bar{D} \rangle_{(D \otimes D)^{-1}}},
$$

(32a)

$$
x_o^{D05} = \frac{\langle \bar{R}, D \rangle_{(D \otimes D)^{-1}}}{\langle \bar{R}, \bar{D} \rangle_{(D \otimes D)^{-1}}},
$$

(32b)

These scaling parameters are given by the ratio of the inner products of $HH^T$ or $\bar{R}$ with the observed and modelled innovation covariance. The only difference between the two methods is their choice of weights for the inner product. The cost functions $J_b$ and $J_o$ evaluated at the BLUE analysis, as given in Equations (2a) and (2b), can also be expressed in a similar manner as

$$
2J_b(x_b) = \langle HH^T, D \rangle_{(D \otimes D)^{-1}},
$$

(33a)

$$
2J_o(x_o) = \langle \bar{R}, D \rangle_{(D \otimes D)^{-1}}.
$$

(33b)

We can see that $2J_b(x_b)$ and $2J_o(x_o)$ are given by the Frobenius inner products of the observed innovation covariance $D$ with $HH^T$ and $\bar{R}$, respectively, weighted by $(D \otimes D)^{-1}$. The $\chi^2$ for the innovation statistics can be expressed as

$$
\chi^2 = d^T \bar{D}^{-1} d = \text{Tr} \left[ \bar{D} \bar{D}^{-1} \right] = \langle \bar{D}, D \rangle_{(D \otimes D)^{-1}}.
$$

(34)

All of these expressions involve inner products between covariance, either weighted by $(D \otimes D)^{-1}$ or $(D \otimes I)^{-1}$. As such, inner products between covariances play a central role in these methods.

In this section, we will develop an intuitive understanding of the inner product between covariances. With a matrix norm and angle defined via the Frobenius inner product, we can interpret the covariances geometrically. This geometric interpretation is not merely by analogy, as each covariance matrix can be converted into a vector by
use of the vectorization operator. As weighting these inner products by \((D \otimes D)^{-1}\) yields unitless products, we will illustrate these products using this weighting, but note that similar results are obtained when weighting by \((D \otimes I)^{-1}\) instead.

### 8.2 Spectral distinctiveness of analysis filters

As a first step, we consider inner products where both of the operands are modelled quantities. The inner products \(\langle BH^T D, D \rangle_{(D \otimes D)^{-1}}\) and \(\langle \bar{R}, D \rangle_{(D \otimes D)^{-1}}\), which appear in Equations \((31a)\) and \((31b)\) can be decomposed as

\[
\langle BH^T D, D \rangle_{(D \otimes D)^{-1}} = ||BH^T||^2_{(D \otimes D)^{-1}} + \langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}},
\]

\[
(35a)
\]

\[
\langle \bar{R}, D \rangle_{(D \otimes D)^{-1}} = ||\bar{R}||^2_{(D \otimes D)^{-1}} + \langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}},
\]

\[
(35b)
\]

which involve the norms of \(BH^T\) and \(\bar{R}\) and the inner product between them. Using the relations in Equations \((5a)\) and \((5b)\), these factors can be written as

\[
\langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}} = \text{Tr} \left[ BH^T D^{-1} R D^{-1} \right] = \text{Tr} \left[ F (I - \bar{F}) \right] = \sum_i \frac{\phi_i}{(1 + \phi_i)^2},
\]

\[
(36a)
\]

\[
||BH^T||^2_{(D \otimes D)^{-1}} = \text{Tr} \left[ BH^T D^{-1} BH^T D^{-1} \right] = \text{Tr} \left[ \Phi \right] = \sum_i \frac{1}{(1 + \phi_i)^2},
\]

\[
(36b)
\]

\[
||\bar{R}||^2_{(D \otimes D)^{-1}} = \text{Tr} \left[ R D^{-1} R D^{-1} \right] = \text{Tr} \left[ (I - \bar{F}) (I - \bar{F}) \right] = \sum_i \frac{\phi_i^2}{(1 + \phi_i)^2},
\]

\[
(36c)
\]

where \(\phi_i\) is the \(i\)th diagonal element of the matrix \(\Phi\). From Equations \((36b)\) and \((36c)\), we see that \(||BH^T||^2_{(D \otimes D)^{-1}}\) and \(||\bar{R}||^2_{(D \otimes D)^{-1}}\) are equal to the sum of the square of the filters \(\bar{F}\) and \(I - \bar{F}\), or \((1 + \phi_i)^{-1}\) and \(\phi_i(1 + \phi_i)^{-1}\) in spectral space, respectively. \(\langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}}\) measures the overlap between \(\bar{F}\) and \(I - \bar{F}\), which is the sum of the product of \((1 + \phi_i)^{-1}\) and \(\phi_i(1 + \phi_i)^{-1}\), and quantifies how distinguishable \(BH^T\) is from \(\bar{R}\) (or \(\bar{F}\) is from \(I - \bar{F}\)) based on their spectra.

We can further quantify how spectrally distinct the filters \(\bar{F}\) and \(I - \bar{F}\) are from one another by using Equation \((12)\) to define the angle \(\theta_{\bar{R}, \bar{F}}\) between \(BH^T\) and \(\bar{R}\) weighted by \((D \otimes D)^{-1}\) as

\[
\cos(\theta_{\bar{R}, \bar{F}}) = \frac{\langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}}}{||BH^T||_{(D \otimes D)^{-1}} \cdot ||\bar{R}||_{(D \otimes D)^{-1}}},
\]

\[
(37)
\]

In this nomenclature, factors of \(H\) in subscript of \(\theta_{\bar{R}, \bar{F}}\) are assumed but not written explicitly in order to simplify notation. Additionally, a weighting of \((D \otimes D)^{-1}\) is assumed for \(\theta_{\bar{R}, \bar{F}}\) unless stated otherwise. Accordingly, \(\cos(\theta_{\bar{R}, \bar{F}})\) is the ratio of the filter overlap term to the product of the root-sum-square of each filter.

As an illustration, consider the case where the spectra of \(BH^T\) and \(\bar{R}\) are such that, for the wavenumbers where the spectrum of \(BH^T\) is significant, the spectrum of \(\bar{R}\) is negligible, and at the wavenumbers where the spectrum of \(\bar{R}\) is significant the spectrum of \(BH^T\) is negligible. We can say that the spectral overlap between \(BH^T\) and \(\bar{R}\) is negligible in this situation (lower plot of Figure 1d for a case where the spectral overlap is relatively small). In this case, at most wavenumbers, either the spectrum of the filter \(\bar{F}\) is close to unity and the spectrum of the filter \(I - \bar{F}\) is close to zero (i.e., \((1 + \phi_i)^{-1} \approx 1\) and \(\phi_i(1 + \phi_i)^{-1} \approx 0\), or vice versa. This implies that the filters \(\bar{F}\) and \(I - \bar{F}\) can relatively unambiguously attribute a particular wavenumber to either the background or the observations in this situation. Consequently, at almost all wavenumbers, the product between \((1 + \phi_i)^{-1}\) and \(\phi_i(1 + \phi_i)^{-1}\) would be negligible and \(\langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}}\), being the sum of this product over all wavenumbers, would be small. In the limit of no overlap between the spectra of \(BH^T\) and \(\bar{R}\), as the filter spectra \((1 + \phi_i)^{-1}\) and \(\phi_i(1 + \phi_i)^{-1}\) would be equal to either zero or one, \(||BH^T||^2_{(D \otimes D)^{-1}}\) and \(||\bar{R}||^2_{(D \otimes D)^{-1}}\) would be equal to the number of non-zero modes in \(BH^T\) and \(\bar{R}\), respectively. In this case, Equation \((37)\) implies that \(\theta_{\bar{R}, \bar{F}}\) would be close to \(90^\circ\), and so we can consider \(BH^T\) and \(\bar{R}\) to be (nearly) orthogonal to one another. Notice from Equation \((36a)\) that, if \(F\) is a projection operator (i.e., \(F^2 = F\)), then \(BH^T\) and \(\bar{R}\) will be orthogonal to one another.

On the other extreme, if the spectral structure of \(BH^T\) and \(\bar{R}\) are similar to one another, the spectra of \(BH^T\) and \(\bar{R}\) will be significant (or insignificant) at the same wavenumbers. In other words, the two spectra will overlap considerably. In this case, the filters \(\bar{F}\) and \(I - \bar{F}\) will have similar spectra and consequently there will be few wavenumbers that could be used to distinguish between the two error covariances. In this case, at most wavelengths the product between \((1 + \phi_i)^{-1}\) and \(\phi_i(1 + \phi_i)^{-1}\) would be significant, leading to a non-negligible value of \(\langle BH^T, \bar{R} \rangle_{(D \otimes D)^{-1}}\).

We have seen in previous sections that both the DI01 and D05 algorithms can be viewed as fitting procedures for the filtered innovation covariances \(D_R\) and \(D_{\bar{R}}\). As will be demonstrated in later sections, the behaviour of these algorithms will depend on how well the filters attribute a particular wavenumber of the innovation covariance to either the background or the observations. The angle \(\theta_{\bar{R}, \bar{F}}\) is one possible measure to quantify how well the filters \(\bar{F}\)
and $I - \hat{F}$ isolate the background and observational contributions to the innovation covariance. Another measure examines the differences $BH^{T} - D_{B}$ and $\hat{R} - D_{R}$, or, in other words, the differences between the full error covariance matrix (either the background or observation) and the appropriately filtered innovation covariance. It can be easily shown that $BH^{T} - D_{B} = R - D_{R} = \hat{F}D(I - \hat{F})^{T} = BH^{T}D^{-1}\hat{R}$, which has a norm give by

$$
||BH^{T} - D_{B}||_{(D \otimes D)^{-1}}^{2} = ||\hat{R} - D_{R}||_{(D \otimes D)^{-1}}^{2} = \langle BH^{T}D^{-1}\hat{R} \rangle_{(D \otimes D)^{-1}}^{2} = Tr[\hat{F}(I - \hat{F})\hat{F}(I - \hat{F})]. \tag{38}
$$

Comparing this relation with Equation (36a), we see that the square of this norm is similar to $\langle BH^{T}\hat{R} \rangle_{(D \otimes D)^{-1}}$ but contains two factors of $\hat{F}(I - \hat{F})$ within the trace instead of one. This norm gives a slightly different measure for the degree of spectral overlap between the modes of $BH^{T}$ and $\hat{R}$. If $BH^{T}$ and $\hat{R}$ have no non-zero modes in common (so that $\theta_{B,R} = 90^{\circ}$), then the filters $\hat{F}$ and $I - \hat{F}$ can perfectly divide the innovation covariance into parts associated with either $BH^{T}$ or $\hat{R}$, in which case we would have $D_{B} = BH^{T}$ and $D_{R} = \hat{R}$. The larger the overlap between the modes of $BH^{T}$ and $\hat{R}$, the further $D_{B}$ is from $BH^{T}$ and $D_{R}$ is from $\hat{R}$.

Lastly, we note that, as these inner products have been weighted by $(I \otimes D)^{-1}$, $||D||_{(D \otimes D)^{-1}}^{2}$ is simply equal to the total number of observations $N_{obs}$. The law of cosines as applied to the triangle formed by $BH^{T}$, $\hat{R}$, and $D$ yields

$$
||D||_{(D \otimes D)^{-1}}^{2} = ||BH^{T}||_{(D \otimes D)^{-1}}^{2} + ||\hat{R}||_{(D \otimes D)^{-1}}^{2} + 2||BH^{T}||_{(D \otimes D)^{-1}} \cdot ||\hat{R}||_{(D \otimes D)^{-1}} \cos(\theta_{B,R}). \tag{39}
$$

Therefore, for a constant number of observations, $||BH^{T}||_{(D \otimes D)^{-1}}$, $||\hat{R}||_{(D \otimes D)^{-1}}$, and $\theta_{B,R}$ are not all independent of one another.

### 8.3 Example: 1D periodic domain

We now illustrate the inner products introduced in the last section with the example of homogeneous covariances defined on a 1D periodic domain, though we emphasize that neither of these properties have been assumed in the above derivations. We also assume that the observations measure the same quantity as the model field at the equally-spaced model grid locations, so that $H = I$. In this example, the simultaneous spectral decomposition of $BH^{T}$ and $\hat{R}$ coincides with the Fourier decomposition. Observation errors are assumed to be spatially uncorrelated, while the background correlations are modelled using a Gaussian function with correlation length $L_{b}$. The modelled homogeneous background- and observation-error variances are denoted by $\tilde{\sigma}_{b}^{2}$ and $\tilde{\sigma}_{o}^{2}$, respectively, and the domain of length is denoted by $L$.

The modelled background- and observation-error covariances for the 1D periodic example is shown in spectral space in Figure 1 for four different choices of parameters. In each panel, the top plots show the background- and observation-error covariances, while the bottom plots show the filter functions $(1 + \phi)^{-1}$ and $\phi(1 + \phi)^{-1}$ (corresponding to the filters $\hat{F}$ and $I - \hat{F}$, respectively) as well as the product between these filters. As the observation errors were assumed to be uncorrelated in space, its spectrum is flat. It is clear that $(1 + \phi)^{-1}$ and $\phi(1 + \phi)^{-1}$ act as a low-pass and a high-pass filter, respectively. As seen in the different panels of Figure 1, the overlapping region of the filters is larger for smaller values of the background correlation length.

The values for the norms of $BH^{T}$ and $\hat{R}$, as well as the angle $\theta_{B,R}$, are shown in each panel. From Equations (36b) and (36c), $||BH^{T}||_{(D \otimes D)^{-1}}^{2}$ and $||\hat{R}||_{(D \otimes D)^{-1}}^{2}$ are given by the sum of the squares under the $(1 + \phi)^{-1}$ and $\phi(1 + \phi)^{-1}$ curves, respectively. From Equation (36a), $\langle BH^{T}\hat{R} \rangle_{(D \otimes D)^{-1}}$ is the sum under the $\phi(1 + \phi)^{-1}$ curve. Comparing Figures 1a and b, the filter $\hat{F}$ in (a) filters out more wavelengths than that chosen for (b), which is reflected in the norm of $BH^{T}$ having a smaller value in (a) than in (b). The filters $\hat{F}$ and $I - \hat{F}$ significantly overlap over more wavenumbers in (b) than in (a), and accordingly the value of $\theta_{B,R}$ is smaller in (b) than in (a). In (b), the value of $\tilde{\sigma}_{b}^{2}/\tilde{\sigma}_{o}^{2}$ was changed from its value in (a) so that the values of the lowest mode in $BH^{T}$ were the same in both panels. To examine changes in length-scales and variances separately, (c) shows the case where the background correlation length is shorter than in (a) but the variances are unaltered, while (d) shows the case where the background correlation length has the same value as in (a) but with different variance values. From these comparisons, we see that, in general, changing either the correlation length or variances changes both the norms of $BH^{T}$ and $\hat{R}$ as well as the angle between them.

Figure 2 shows $\theta_{B,R}$, $||BH^{T}||_{(D \otimes D)^{-1}}^{2}$, and $||\hat{R}||_{(D \otimes D)^{-1}}^{2}$ as a function of the background correlation length $L_{b}$ for a variety of different observation spacings $\Delta x$. In this example, we set the background-and observation-error variances equal to one another and set the domain length to $L = 40,000$ km, which roughly coincides with the Earth's circumference. As $L_{b}$ increases, the background correlations become more spectrally distinct from the uncorrelated observation uncertainties and accordingly $\theta_{B,R}$ moves closer to $90^{\circ}$. Increasing $L_{b}$ also decreases the spectral extent of the filter $\hat{F}$ (compare the bottom plots of Figures 1a and c), and so decreases the value of
$|\mathbf{H}\mathbf{B}\mathbf{H}^T|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$, as seen in Figure 2b. Conversely, increasing $L_b$ increases the spectral extent of the filter $\mathbf{I} - \mathbf{F}$, so the values of $||\mathbf{R}||_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ in Figure 2b increase with $L_b$.

Most practical data assimilation applications involve making observations at discrete locations, in which case we are able to sample the covariances only at these locations. Figure 2 demonstrates that the observation spacing has a significant effect on the spectral overlap between $\mathbf{H}\mathbf{B}\mathbf{H}^T$ and $\mathbf{R}$. If a correlation function is sampled at discrete locations and the spaces between samples are greatly increased, then the sampled correlation function will begin to look more like a delta function, assuming the continuous correlation function that is being sampled approaches zero at infinite distance. Conversely, if the spaces between samples are greatly decreased, the correlation values between adjacent locations do not change as much, so the sampled correlation function starts to look more flat. Thus if the observation errors are uncorrelated in space, or have a much smaller correlation length than the background errors, increasing the observation spacing increases the spectral overlap between $\mathbf{H}\mathbf{B}\mathbf{H}^T$ and $\mathbf{R}$ as the two sampled correlations becomes less distinct from one another. This is reflected in Figure 2, where we see that for a fixed value of $L_b$, $\theta_{\mathbf{R},\mathbf{R}}$ decreases as $\Delta x$ increases. But as the observations are equally spaced in this example, for a set value of $L$, changing $\Delta x$ also changes the number of observation locations. Figure 3 examines this issue by plotting $\theta_{\mathbf{R},\mathbf{R}}$ as a function of observation spacing $\Delta x$, but holds the number of observation locations constant, so that $L$ varies with $\Delta x$. Here too we see that $\theta_{\mathbf{R},\mathbf{R}}$ decreases as the observation spacing increases. Returning to Figure 2b, we see that increasing the density of observations increases both $|\mathbf{H}\mathbf{B}\mathbf{H}^T|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ and $||\mathbf{R}||_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ for a fixed value of $L_b$. Both $|\mathbf{H}\mathbf{B}\mathbf{H}^T|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2$ and $||\mathbf{R}||_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2$ can be interpreted as an integral over the square of their respective filters (the squared values under the blue and red curves in the bottom plots of Figure 1). As increasing the observation density increases the number of modes in observation space, in general we expect this to increase the numerical values of these integrals, which would increase the values of both $|\mathbf{H}\mathbf{B}\mathbf{H}^T|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ and $||\mathbf{R}||_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$.

Up to this point, all examples examined have had uncorrelated observation errors. Figure 4 shows plots of the covariances and filters with the same model parameters as those in Figure 1a, but with observation errors correlated in space, modelled by a Gaussian with correlation length $L_o$. The observation-error correlation length in Figure 4b is five times larger than its value in Figure 4a. In contrast to the flat observation-error spectra in Figure 1, those in Figure 4 decrease with wavenumber. Although the observation-error covariance spectra in Figure 4 decreases significantly at high wavenumbers, in these examples the observation-error spectra are still significantly larger than the background-error covariance spectrum at these wavenumbers. Consequently, the spectra for the filters $\mathbf{F}$ and $\mathbf{I} - \mathbf{F}$ are nearly the same as in
Figure 4 Modelled background- and observation-error covariances as a function of wavenumber for cases with spatially correlated observation errors. In each panel, the top plots show the covariances and the bottom plots show the filter functions. Covariances are homogeneous and defined on a 1D periodic domain with 1001 equally spaced grid points. Both background and observation correlations are modelled using a Gaussian with correlation lengths \( L_b \) and \( L_o \), respectively. Variances are homogeneous with \( \overline{\sigma}^2 = \overline{\sigma}^2_o \). The units of the covariances are arbitrary.

Figure 1a at high wavenumbers. However, differences can be seen at low wavenumbers. Comparing Figures 4a, b to Figure 1a, we can see that the observation-error covariance spectrum is closer to the background-error covariance spectrum at low wavenumbers for the spatially correlated cases as compared to the uncorrelated case. As such, it is more difficult to distinguish between the background- and observation-error covariance at these wavenumbers, as seen by the decrease of the spectrum of \( \Phi \) and increase of the spectrum of \( \mathbf{I} - \Phi \) with increasing observation-error correlation length at low wavenumbers. Comparing Figure 1a with Figure 4, we see that the value of \( \theta_{b,R} \) decreases with increasing observation-error correlation length, which in these examples is primarily due to the differences in spectra at low wavenumbers. This is further illustrated in Figure 5, which shows the angle \( \theta_{b,R} \) as a function of observation-error correlation length for three different values of the background-error correlation length. Figure 5 clearly shows that \( \theta_{b,R} \) decreases as the observation-error correlation length approaches the value of the background-error correlation length.

The examples chosen for Figures 4 and 5 all have \( L_b > L_o \). For cases with \( L_b < L_o \), as for example atmospheric motion vectors (e.g., Cordoba et al. (2017)), we can use the same examples that were used for cases with \( L_b > L_o \) but with \( \mathbf{HBH}^T \) and \( \mathbf{R} \) interchanged. From Equations (36a) and (37), we note that interchanging \( \mathbf{HBH}^T \) and \( \mathbf{R} \) does not alter \( \langle \mathbf{HBH}^T, \mathbf{R} \rangle/\langle \mathbf{D} \otimes \mathbf{D} \rangle \), or \( \theta_{b,R} \), while from Equations (36b) and (36c) the values of \( ||\mathbf{HBH}^T||_{\mathbf{D} \otimes \mathbf{D}^{-1}}^2 \) and \( ||\mathbf{R}||_{\mathbf{D} \otimes \mathbf{D}^{-1}}^2 \) will be switched. Therefore, the plots in Figure 5 will be identical to the same plots with the values of \( L_b \) and \( L_o \) switched, provided that we switch the labels for \( L_b \) and \( L_o \), and similarly with Figure 4 additionally switching the labels for \( \mathbf{HBH}^T \) and \( \tilde{\mathbf{R}} \), as well as the labels for \((1 + \phi)^{-1}\) and \(\phi(1 + \phi)^{-1}\).

We saw in Sections 5.2 and 6.2 that the filtered innovation covariances play a central role in the both the DI01 and D05 methods, as it enables the algorithm to attribute parts of the innovation covariance to either the background- or observation-error covariance. Figure 6 illustrates the filtering of the innovation covariance by showing \( \mathbf{HBH}^T \) and \( \mathbf{D}_B \) for four different choices of parameters as a function of separation distance \( x \). Figure 6a, b have \( \overline{\sigma}^2_o = \overline{\sigma}^2_b \), while Figure 6c, d have values of \( \overline{\sigma}^2_o \) decreased to half those in (a, b). Figure 6a, c use a background correlation length of \( L_b = 1,000 \text{km} \), while (b, d) use \( L_b = 100 \text{km} \). The values for \( ||\mathbf{HBH}^T - \mathbf{D}_B||_{\mathbf{D} \otimes \mathbf{D}^{-1}} \) and \( \theta_{b,R} \) are also displayed for each case. Figure 6b, d use a smaller value of \( L_b \) than in (a, c), and correspondingly (b, d)
have smaller $\theta_{B,R}$ values and larger $\| \mathbf{H} \mathbf{B}^T - \mathbf{D}_B \|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ values than in (a, c). Now comparing (a, b) with (c, d) respectively, as $\sigma_o^2$ decreases, the observational uncertainty contributes less to the innovation covariance and so the observational covariances misattributed to the background will be comparatively smaller as well. This is evident by comparing (a, b) with (c, d) respectively, where (c, d) have larger values for $\theta_{B,R}$ and smaller values for $\| \mathbf{H} \mathbf{B}^T - \mathbf{D}_B \|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ than (a, b), respectively. These examples illustrate that the more spectrally distinct $\mathbf{H} \mathbf{B}^T$ and $\mathbf{R}$ are to one another, the closer $\mathbf{D}_B$ is to $\mathbf{H} \mathbf{B}^T$ (and $\mathbf{R}$ is to $\mathbf{R}$).

8.4 Limiting cases revisited

One of the fundamental limitations of error covariance estimation is related to the lack of identifiability between background and observational covariances when correlation length-scales are identical, as pointed out by Dee and Da Silva (1999), Dee et al. (1999) and Chapnik et al. (2004). With this in mind, Chapnik et al. (2004) considered the performance of the DI01 algorithm in two different limiting cases. One case has the background- and observation-error correlation lengths equal, and a contrasting case where the background-error correlation length is taken to be infinite. We now revisit these cases to gain further intuitive insight into the inner products between covariances.

In the first case, an infinite background-error correlation length implies that all background-error correlations are equal to unity, which is the limiting case for situations where the background-error correlation length is much larger than the length of the domain. In contrast, the observation errors are taken as uncorrelated. In this case, $\mathbf{H} \mathbf{B}^T$ would be a rank-1 matrix with a single non-zero spectral value while $\mathbf{R}$ would have a uniform spectrum. Consequently, as $\phi_i$ is the equal to the observation-error covariance spectrum divided by the background-error covariance spectrum at wavenumber $i$, the value of $\phi_i$ would be infinite at every wavenumber other than at the lowest wavenumber. Examining the sums in Equations (36a) and (36b), only the first terms in the sums would be non-zero. The spectral overlap between the filters $\tilde{\mathbf{F}}$ and $\mathbf{I} - \tilde{\mathbf{F}}$ in this case is confined to solely the lowest wavenumber. However, the sum in Equation (36c) taken over all terms except the first would be equal to $N_{\text{obs}} - 1$. Therefore, if $N_{\text{obs}}$ is very large, $\| \mathbf{R} \|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2$ would be much larger than both $\langle \mathbf{H} \mathbf{B}^T, \mathbf{R} \rangle_{(\mathbf{D} \otimes \mathbf{D})^{-1}}$ and $\| \mathbf{H} \mathbf{B}^T \|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2$, in which case, from Equation (37), $\cos(\theta_{B,R})$ would be small and $\theta_{B,R}$ would be close to 90°. As $\cos(\theta_{B,R})$ continues to decrease with increasing values of $N_{\text{obs}}$, $\mathbf{H} \mathbf{B}^T$ and $\mathbf{R}$ are asymptotically orthogonal to one another (with respect to the weighting $(\mathbf{D} \otimes \mathbf{D})^{-1}$) in this case. As $\mathbf{H} \mathbf{B}^T$ and $\mathbf{R}$ are nearly orthogonal in this case, the analysis filter can attribute certain modes to either the background or observations with little error.

The second case is where the background- and observation-error covariances have the exact same
correlation and variance structure, so that $\mathbf{HBH}^T$ and $\mathbf{R}$ are equal up to an overall scaling factor. It can easily be verified that all values of $\phi_i$ are equal to one another in this case, which implies that the filters $\mathbf{F}$ and $\mathbf{I} - \mathbf{F}$ both have flat spectra. As the relative importance of $\mathbf{HBH}^T$ compared to $\mathbf{R}$ is the same for each wavenumber, there is no way to use the spectral features of the innovation covariance to attribute a particular part to either the background or the observations. Using Equations (36a) to (36c) for this example gives $\cos(\theta_{\mathbf{R},\mathbf{R}}) = 1$, so that $\theta_{\mathbf{R},\mathbf{R}} = 0^\circ$ and $\mathbf{HBH}^T$ and $\mathbf{R}$ can be considered to be parallel to one another (with respect to the weighting $(\mathbf{D} \otimes \mathbf{D})^{-1}$). Consequently, in this case, the analysis filter is ineffective at attributing modes to either the background or observations.

Note that, in the second example, both the background- and observation-error covariances have the same covariance structure, not just the same correlation structure. As the rightmost expressions in Equations (36a) to (36c) involve the spectra of the full covariance, $\theta_{\mathbf{R},\mathbf{R}}$ may differ from $0^\circ$ in cases where the correlations of $\mathbf{HBH}^T$ and $\mathbf{R}$ are the same but the structure of their variances differ.

8.5 Geometric interpretation of the DI01 and D05 algorithms

As observed in Desroziérs (2012) and Ménard (2016), convergence of the D05 algorithm is slow when the background- and observation-error correlations have similar length-scales. As the difference in correlation structure between the two error covariances can be quantified using inner products of the form $(\mathbf{HBH}^T, \mathbf{R})$ (and the corresponding angle $\theta_{\mathbf{R},\mathbf{R}}$), these quantities can be used to quantify the rate of convergence for the DI01 and D05 algorithms. Additionally, since this has been quantified using an inner product, we can make geometric interpretations of the DI01 and D05 algorithms.

The variance scaling parameters for the DI01 method, as given in Equations (31a) and (31b), can be expressed in terms of the angle $\theta_{\mathbf{R},\mathbf{D}}$ between $\mathbf{HBH}^T$ and $\mathbf{D}$ and $\theta_{\mathbf{R},\mathbf{D}}$ between $\mathbf{HBH}^T$ and $\mathbf{D}$ as

$$s_{b}^{\text{DI01}} = \frac{\|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}} \cos(\theta_{\mathbf{R},\mathbf{D}})}{\|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}} \cos(\theta_{\mathbf{R},\mathbf{D}})}, \quad (40a)$$

$$s_{o}^{\text{DI01}} = \frac{\|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}} \cos(\theta_{\mathbf{R},\mathbf{D}})}{\|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}} \cos(\theta_{\mathbf{R},\mathbf{D}})}. \quad (40b)$$

Written in this manner, it is evident that $s_{b}^{\text{DI01}}$ is the ratio of the projection of $\mathbf{D}$ onto $\mathbf{HBH}^T$ to the projection of $\mathbf{D}$ onto $\mathbf{HBH}^T$. This scaling is depicted in Figure 7, which represents the inner product space defined by the Frobenius inner product by $(\mathbf{D} \otimes \mathbf{D})^{-1}$, so that matrices are represented by vectors with lengths given by their

![Figure 7: Geometrical representation of the DI01 and D05 algorithms.](image)

8.6 Geometric interpretation of the $\chi^2$ diagnostic

We saw in Equation (34) that the $\chi^2$ diagnostic could be expressed as the inner product between the observed ($\mathbf{D}$) and modelled ($\mathbf{D}$) innovation covariances weighted by $(\mathbf{D} \otimes \mathbf{D})^{-1}$. As the expectation value of $\chi^2$ is equal to the number of observations $N_{\text{obs}}$ when the modelled and observed innovation covariances are consistent with one another, the $\chi^2$ diagnostic relation compares the observed $\chi^2$ value to $N_{\text{obs}}$. The $\chi^2$ diagnostic relation can be written in terms of the Frobenius inner product as

$$\langle \mathbf{D}, \mathbf{D} \rangle_{(\mathbf{D} \otimes \mathbf{D})^{-1}} = \|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 = N_{\text{obs}}. \quad (41)$$

Complete innovation consistency implies that $\mathbf{D} = \mathbf{D}$ for the full covariance matrices, while the $\chi^2$ diagnostic looks solely at the quantity $\langle \mathbf{D}, \mathbf{D} \rangle_{(\mathbf{D} \otimes \mathbf{D})^{-1}} = \mathbf{d}^T \mathbf{D}^{-1} \mathbf{d}$. As such, when the $\chi^2$ diagnostic is satisfied, the modelled and observed innovation covariances will still in general differ from one another. The degree of difference between the observed and prescribed innovation covariance can be quantified by

$$\|\mathbf{D} - \mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 = \|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 - 2 \langle \mathbf{D}, \mathbf{D} \rangle_{(\mathbf{D} \otimes \mathbf{D})^{-1}} + \|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 = \|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 - 2\chi^2 + N_{\text{obs}}. \quad (42)$$

If the $\chi^2$ diagnostic is satisfied, then Equation (42), when rearranged, simplifies to

$$\|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 + \|\mathbf{D} - \mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2 = \|\mathbf{D}\|_{(\mathbf{D} \otimes \mathbf{D})^{-1}}^2. \quad (43)$$
Equation (43) implies that when the \( \chi^2 \) diagnostic is satisfied, the norm of \( \mathbf{D} - \bar{\mathbf{D}} \) can be found via the Pythagorean theorem. This situation is represented in Figure 8. Therefore, the difference between the squared norms of \( \mathbf{D} \) and \( \bar{\mathbf{D}} \) quantifies the level of difference between the observed and modelled innovation covariances that remains after the \( \chi^2 \) diagnostic has been satisfied.

9 | Statistical Properties of a Priori and a Posteriori Methods

Up to this point, the comparison between the DI01, D05, and HL methods has been conceptual. In this section, we begin to examine the statistical properties of these methods by comparing the expectation values and variances of the scaling parameters derived by each method. This will be done in the idealized case where the structures of \( \mathbf{B} \) and \( \mathbf{R} \) are correctly specified and vary from the truth only by global scaling factors. In other words, in this case the true covariance matrices would be given by \( \mathbf{B}^t = s^t_b \mathbf{B} \) and \( \mathbf{R}^t = s^t_o \mathbf{R} \), where \( s^t_b \) and \( s^t_o \) are the true variance scalings required for our modelled covariances. The expressions in this section may be extended to include the effects of mis-specifications of the error correlations, but are not included here for brevity.

We begin by examining the expectation values of the DI01 scaling parameters. The expectation values can be found by the use of the relation in Equation (B11) with Equations (15a) and (15b), which with some simple algebraic manipulation yields

\[
\frac{E[s^t_{D101}]}{s^t_b} = 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\langle \mathbf{H}^T \mathbf{B}^T, \mathbf{R} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}{\langle \mathbf{H}^T \mathbf{D}^T, \bar{\mathbf{D}} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}
= 1 + \left( \frac{s^t_o}{s^t_o} - 1 \right) \times \left[ 1 + \frac{\mathbf{H}^T \mathbf{B}^T}{\mathbf{R}^T \mathbf{D} \bar{\mathbf{D}}^{-1}} \frac{1}{\cos(\theta_{R,R})} \right]^{-1},
\]

(44a)

\[
\frac{E[s^t_{D01}]}{s^t_o} = 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\langle \mathbf{H}^T \mathbf{R} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}{\langle \mathbf{R}^T \mathbf{D} \bar{\mathbf{D}}^{-1} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}
= 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \times \left[ 1 + \frac{\mathbf{H}^T \mathbf{R}}{\mathbf{R}^T \mathbf{D} \bar{\mathbf{D}}^{-1}} \frac{1}{\cos(\theta_{R,R})} \right]^{-1},
\]

(44b)

From the first lines of Equations (44a) and (44b), we can see that the expectation values of the scaling parameters are equal to their true values if \( \langle \mathbf{H}^T \mathbf{B}^T, \mathbf{R} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1} \) is equal to zero or, in other words, when the spectral overlap between \( \mathbf{H}^T \mathbf{B} \) and \( \bar{\mathbf{R}} \) is negligible. When \( \mathbf{B} \) and \( \bar{\mathbf{R}} \) differ from their true values by the same scaling factor, so that \( s^t_b = s^t_o \), the expectation values of the scaling parameters will also equal their true values, in which case the updated error statistics do not effect the analysis. From the second line of Equation (44a), we can see that as \( \|\mathbf{H}^T \mathbf{B}^T\|_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}/\|\mathbf{R}\|_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1} \) increases, \( E[s^t_{D101}] \) approaches its true value due to the decreasing influence of \( \mathbf{R} \) on \( \mathbf{B}^t \), and vice versa for \( E[s^t_{D01}] \) as \( \|\mathbf{H}^T \mathbf{B}^T\|_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}/\|\mathbf{R}\|_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1} \) decreases. From the second lines of Equations (44a) and (44b), we can see that the factors multiplying \( s^t_b/s^t_b - 1 \) or \( s^t_b/s^t_b - 1 \) will have the same sign, while \( s^t_b/s^t_o - 1 \) and \( s^t_b/s^t_o - 1 \) will have opposite signs, so if \( E[s^t_{D101}] \) is overestimated, then \( E[s^t_{D01}] \) will be underestimated, and vice versa.

Turning now to the D05 method, using Equation (B11) with Equations (20a) and (20b) yields the expectation values for the D05 variance scaling parameters

\[
\frac{E[s^t_{D05}]}{s^t_b} = 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\langle \mathbf{H}^T \mathbf{B}^T, \bar{\mathbf{R}} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}{\langle \mathbf{H}^T \mathbf{D}^T, \bar{\mathbf{D}} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}
= 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\text{Tr}[\mathbf{H}^T \mathbf{B}^T \bar{\mathbf{R}}]}{\text{Tr}[\mathbf{H}^T \bar{\mathbf{D}}]},
\]

(45a)

\[
\frac{E[s^t_{D05}]}{s^t_o} = 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\langle \mathbf{H}^T \mathbf{R}, \bar{\mathbf{D}} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}{\langle \mathbf{R}, \bar{\mathbf{D}} \rangle_{\mathbf{D} \otimes \bar{\mathbf{D}}}^{-1}}
= 1 + \left( \frac{s^t_b}{s^t_b} - 1 \right) \frac{\text{Tr}[\mathbf{H}^T \mathbf{R} \bar{\mathbf{D}}]}{\text{Tr}[\mathbf{R} \bar{\mathbf{D}}]},
\]

(45b)

As with the DI01 variance scaling parameters, we see that the divergence of the expectation values of the D05 variance scaling parameters from their true values is proportional to terms that measure the spectral overlap between \( \mathbf{H}^T \mathbf{B} \) and \( \bar{\mathbf{R}} \). For the weightings chosen in the D05 method, the spectral overlap term is given by \( \text{Tr}[\mathbf{H}^T \mathbf{B} - \mathbf{R}] \).
\[ \tilde{D}_b = \text{Tr}[\tilde{R} - \tilde{D}_b]. \]

As for the DI01 estimates, if one scaling parameter is overestimated, then the other will be underestimated.

In this example, we have assumed that the modelled error covariances only differ from their true values by a scaling factor. As the HL method assumes spatially uncorrelated observations, this example implies that this assumption is satisfied. In this case, the expectation values of the HL variance scaling parameters \( s_b^{\text{HL}} \) and \( s_o^{\text{HL}} \) are equal to their true values,

\[
E[s_b^{\text{HL}}] = s_b, \quad (46a) \\
E[s_o^{\text{HL}}] = s_o. \quad (46b)
\]

although we emphasize that this only applies to cases where the true observation-error covariances are uncorrelated. If the true observation uncertainties are spatially correlated, then the relations in Equations (46a) and (46b) would not hold.

For the DI01 and D05 methods, any spectral overlap between the background- and observation-error covariances cause some amount of mis-attribute of the source of the innovation covariance. In contrast, if the observation-error covariance is truly uncorrelated in space, the HL method would correctly attribute all spatial correlations in the innovation covariance to the background. Another way to express this is by noting that the inner product \( \langle V o D, I o D \rangle \) vanishes for any diagonal weighting matrix. This inner product serves an analogous role in the HL method to the inner product \( \langle \tilde{H}^T B, \tilde{R} \rangle \) in the DI01 and D05 methods (weighted by the appropriate factor). While the HL method does not make the types of attribution errors as in the DI01 and D05 methods, it is limited by the assumption of spatially uncorrelated observation errors.

Expressions for the variances of the scaling factors for these methods in this example can be found in Appendix D.

### 10 | COMPARISON BETWEEN METHODS

In this section, we compare the scaling parameter estimates made using the DI01, D05, and HL methods for the example of homogeneous covariances defined on a 1D periodic domain, which was introduced in Section 8.3. As before, we restrict our analysis to global scalings of the error variances. We assume that both the modelled and true variances are homogeneous and correlations are modelled by a Gaussian. As in Section 9, we assume that the true and modelled covariances only differ by a scaling factor (Waller et al. (2016a) give a discussion of results for

the D05 method when the true and assumed correlation models differ). In this case, the true scaling parameters are given by \( s_b = (\sigma_b^2)^{1/2} \) and \( s_o = (\sigma_o^2)^{1/2} \), where \( (\sigma_b^2)^2 \) and \( (\sigma_o^2)^2 \) are the true background- and observation-error variances, respectively. Therefore, \( s_b^{4}/s_b^{2} \) is simply given by

\[
\frac{s_b^{4}}{s_b^{2}} = \frac{\sigma_b^4}{\sigma_b^2}, \quad (47)
\]

For the DI01 and D05 methods, results are displayed for a single iteration of the algorithm, as only one iteration is performed in most practical implementations. All plots in this section use a domain length of \( L = 40,000 \text{ km} \).

This section will present results for both the expectation values and the variances of the scaling parameters for each method, both of which were calculated analytically using the relations in Section 9 and Appendix D. While the expressions for the expectation values do not require the innovations to be normally distributed, the analytic formulations for the expectance hold in general if the innovations are normally distributed. As such, we assume in this section that the innovations are normally distributed.

We begin by considering the case where both the modelled and true observation errors are spatially uncorrelated. Figure 9 shows the expectation values and square roots of the variances of the scaling parameters for each method as a function of background correlation length \( \Delta x = 40 \text{ km} \) for two different choices of modelled and true error variances. The left panel show a case where \( s_b^{4}/s_b^{2} = 0.5 \), while the right panel have \( s_b^{4}/s_b^{2} = 0.5 \). The observation spacing is set at \( \Delta x = 40 \text{ km} \) for all panels. For the DI01 and D05 algorithms, the expectation values of the scaling parameters approach their true values as \( \Delta x \) increases since this decreases the spectral overlap between \( \tilde{H}^T B \) and \( \tilde{R} \). As the true observation errors are uncorrelated in this case, the expectation values of the HL variance scaling parameters equal their true values. As noted in Section 9, for the DI01 and D05 methods, if one scaling parameter is overestimated then the other is underestimated, and in this example the parameters over- and underestimated switch from the left to the right panels as the values of \( s_b^{4}/s_b^{2} = 1 \) and \( s_b^{4}/s_b^{2} = 1 \) switch sign. Interestingly, increasing \( \Delta x \) increases the variances of many of the scaling parameters for this range of background correlation lengths. While the expectation values of the HL scaling parameters are equal to their true values in this example, their variances are larger than that estimated by both the DI01 and D05 methods. Although the expectation values for D05 background scaling parameter were closer to the truth than that estimated with the DI01 method in this example, the opposite was true for the observation scaling parameters. The variances of the estimates of the background scaling parameter were larger for the D05 method than the DI01 method,
while the variances of the observation scaling estimates were similar using these two methods.

The expectation values for the HL parameters diverge from their true values if the true observation errors are spatially correlated, a case which is illustrated by Figure 10. In this figure, as in the previous example, the estimates using the DI01 and D05 methods use modelled observation covariances that differ from their true values by only a scaling coefficient and so include these spatial correlations. However, since the HL method cannot handle spatially correlated observation errors, the estimates using the HL method do not include a modelled observation-error correlation. The parameter values used for Figure 10 are the same as used in the left panel of Figure 9, but have observation-error correlations modelled by a Gaussian with correlation length $L_o = 40\ km$. As seen in the figure, the divergence of the expectation values of the HL scaling parameters from their true values are on the same order of magnitude as the estimates from the other methods, but provide overestimates when the other methods underestimate, and vice versa for underestimates.

In Section 8.3, it was shown that the spectral overlap of the background- and observation-error covariances is strongly dependent on the observation spacing. Figure 11 shows the dependency of the estimates of the scaling parameters on the observation spacing $\Delta x$. In this figure, both the modelled and true observation errors are spatially uncorrelated. As the observation spacing increases, the expectation values of the scaling parameters from both the DI01 and D05 methods diverge from their true values. In this range of observation spacings, the variances of the observation-error scalings from all three methods increase with $\Delta x$, while the variances of the background scaling parameters estimated using the HL method increases with $\Delta x$ but decreases for the DI01 and D05 methods.
Figure 11  Expectation values (top plot) and square root of the variances (bottom plot) of scaling parameter estimates as a function of observation spacing Δx. The solid and dashed curves denote the background and observation scaling parameters, respectively. Both the modelled and true observation errors are set to be spatially uncorrelated, while the background correlation length is set to $L_b = 600$ km. The case shown uses $\sigma_b^2 / \sigma_o^2 = 1$, $(\sigma_b^2)^2 / (\sigma_o^2)^2 = 2$, and $L = 40,000$ km.

11  |  Weighing the Cost Functions

We have seen that although the numerics of the DI01 and D05 methods differ, as well as their ease of implementation, both fundamentally solve the same least-squares problem but with different weightings. The choice of weightings and their corresponding variance scaling coefficients for some of the methods examined in this paper are summarized in Table 1. The third and fourth lines of Table 1, which both correspond to weightings for the $\chi^2$ diagnostic, is another example of non-unique weightings for an algorithm. (As a diagnostic, the $\chi^2$ diagnostic technically does not have scaling parameters, but is labelled as such in Table 1 in spite of this as the diagnostic equation is emitted by Equations (23a) and (23b) for those choices of weightings.) While more choices for the weightings are possible, many yield expressions for the scaling parameters that are difficult to evaluate in an operational setting, particularly in the evaluation of the trace in the denominators of Equations (23a) and (23b). As such, the choice of weightings for the DI01 and D05 algorithms are an appropriate choice for practical reasons, not necessarily because they are optimal in some theoretical sense.

As the variance scaling parameters are linear least-squares coefficients, ideally we would expect to use a weighting that corresponds to the covariance of the data being fitted. In our case, this would be the covariance of an estimate of the filtered innovation covariance. In the scalar case, the variance of the sample variance of samples drawn from a normal distribution is proportional to the square of the true variance. Symbolically, if $S^2$ is the unbiased sample variance for a set of independent and identically distributed random scalars drawn from a normal distribution with zero mean and variance $\sigma^2$, then $E[S^2] = \sigma^2$ and $V[S^2] \propto \sigma^4 = E[S^2]^2$, where $V[\cdot]$ is the variance operator. In the multidimensional case, vectors independently drawn from a multivariate normal distribution follow a Wishart distribution, the variances of which are proportional to a sum of products between two covariance elements. In both the scalar and multivariate cases, the variances of the sample (co)variances are proportional to products of the (co)variances themselves. Therefore, ideally the cost functions that fit modelled to observed filtered innovation covariances would use weightings that are proportional to the inverse of products of the same filtered innovation covariances being fit. The weightings in Table 1 are inverses of Kronecker products of either the filtered covariances themselves, covariances that have a similar structure (such as the unfiltered covariances), or the identity. Although in most cases the weightings do not exactly match what would be used in the ideal case (the covariance of the Wishart distribution for $D_B$ or $D_R$), when the spectral overlap between the background- and observation-error covariances is small, the weightings in Table 1 would have a similar structure as the ideal weighting. We briefly note that the inverse of the weightings in Table 1 have the same form as the covariance for a matrix normal distribution for matrix-valued variables (Gupta and Nagar (2018) gives a review of matrix variate distributions).

To illustrate the weightings used for the DI01 and D05 algorithms, we return to the example of a 1D periodic domain introduced in Section 8.3. In this example, the background-error correlations are modelled with a second-order autoregressive (SOAR) model, instead of the Gaussian model, while the observation errors are spatially uncorrelated. The top plot of Figure 12 shows the relative weightings in spectral space used in the cost functions associated with the DI01 and D05 algorithms for this example, where the weightings have been scaled so that weighting of the lowest wavenumber is equal to unity (the absolute scaling of the weights is inconsequential). The bottom plot shows the filter functions for comparison. In this example, the filter $\tilde{F}$ (represented by the $(1 + \phi)^{-1}$ curve in spectral space) dampens high wavenumbers, and
associated with a pair of cost functions that measure the mismatch between parts of the modelled and observed innovation covariances. In order for these methods to be successful, these parts must in some way isolate the contributions of the background and observations to the innovation covariance. The DI01 and D05 algorithms both use the filter that produces the analysis to make this separation and as such are categorized as a posteriori methods, which provides some degree of computational savings when analyses are already available. In contrast, the a priori HL method separates the background and observational contributions to the innovation covariance by attributing all spatial correlations to the background.

Estimates for variance scaling factors from each method are obtained by minimizing their respective cost functions. Since each cost function measures the mismatch between observed and modelled covariances, this minimization is a fitting procedure and the variance scaling factors estimated are best-fit values. While this fitting procedure is done explicitly in the HL method, it is implicit in the DI01 and D05 methods. From a conceptual point of view, the DI01 and D05 algorithms fundamentally only differ by their choice of weightings in the implicit fitting procedure. As such, the DI01 and D05 algorithms belong to a larger class of algorithms which are differentiated by their weightings. However, the choice of weightings has important implications for the algorithm’s numerics and its ability to be implemented in an operational setting.

The degree of success of each method is dependent on how well each method performs the separation between the background and observational contributions to the innovation covariance. For the DI01 and D05 methods, this is dependent on the amount of spectral overlap between the background- and observation-error covariances, which can be quantified by the Frobenius inner product between the two covariances. The norms of the error covariances and the angles between them with respect to the Frobenius inner product effectively summarizes important features of the error statistics and can quantify how effective the DI01 and D05 methods will be. Our results suggest that the angle between the background- and observation-error covariances with respect to the Frobenius inner product can be used as a scalar measure of identifiability between the two covariances. Using the Frobenius inner product in this manner also allows for the covariances to be

### TABLE 1

Cost function weightings and variance scaling parameters for the tuning of the error variances for different methods. See text for details

| Method |  $J_R$ weighting |  $J_R$ weighting |  $s_b$ |  $s_o$ |
|--------|------------------|------------------|-------|-------|
| DI01   | $\{D_R \otimes (H^T B H)^T\}^{-1}$ | $(D_R \otimes R)^{-1}$ | $\frac{\text{Tr}(H^T B D D^{-1})}{\text{Tr}(H^T R D D^{-1})}$ | $\frac{\text{Tr}(R D D^{-1})}{\text{Tr}(R D D^{-1})}$ |
| D05    | $(B_R \otimes I)^{-1}$ | $(B_R \otimes I)^{-1}$ | $\text{Tr}(B D D^{-1})$ | $\text{Tr}(B D D^{-1})$ |
| $\chi^2$ | $\{D_R \otimes D_R\}^{-1}$ | $(R_R \otimes R)^{-1}$ | $\text{Tr}(D D^{-1})$ | $\text{Tr}(D D^{-1})$ |

#### FIGURE 12

Relative weightings used in the cost functions associated with the DI01 and D05 algorithms in spectral space for the case of homogeneous covariances defined on a 1D periodic domain. Relative weightings are shown in the top plot, while the filter functions for this case are shown in the bottom plot. In the top plot, weightings used for $I_R^{\text{DI01}}$ and $I_R^{\text{D05}}$ are shown by the solid curves, while the dashed curves show the weightings used for $I_R^{\text{HL}}$ and $I_R^{\text{HL}}$. The observation errors are spatially uncorrelated and a SOAR model is used for the background-error correlations. The model parameters used are $L = 40,000 \text{ km}$, $L_b = 1,000 \text{ km}$, $\Delta x = 40 \text{ km}$, and $\sigma_b^2/\sigma_o^2 = 2$

since the square roots of the $s_b$ weightings are roughly inversely proportional to the innovation covariance filtered by $\tilde{F}$, modes at high wavenumbers are given much higher weights than at lower wavenumbers when determining $s_b$. The same is true for the $s_o$ weightings at lower wavenumbers, where the filter $I - \tilde{F}$ has low amplitudes and the weightings are large compared to those at the higher wavenumbers. This is evident for both the DI01 and D05 algorithms.

### 12 | CONCLUSIONS

We have shown that for the case of estimating error variances, the DI01, D05, and HL methods can all be

---

**TABLE 1**

| Method |  $J_R$ weighting |  $J_R$ weighting |  $s_b$ |  $s_o$ |
|--------|------------------|------------------|-------|-------|
| DI01   | $\{D_R \otimes (H^T B H)^T\}^{-1}$ | $(D_R \otimes R)^{-1}$ | $\frac{\text{Tr}(H^T B D D^{-1})}{\text{Tr}(H^T R D D^{-1})}$ | $\frac{\text{Tr}(R D D^{-1})}{\text{Tr}(R D D^{-1})}$ |
| D05    | $(B_R \otimes I)^{-1}$ | $(B_R \otimes I)^{-1}$ | $\text{Tr}(B D D^{-1})$ | $\text{Tr}(B D D^{-1})$ |
| $\chi^2$ | $\{D_R \otimes D_R\}^{-1}$ | $(R_R \otimes R)^{-1}$ | $\text{Tr}(D D^{-1})$ | $\text{Tr}(D D^{-1})$ |

#### FIGURE 12

Relative weightings used in the cost functions associated with the DI01 and D05 algorithms in spectral space for the case of homogeneous covariances defined on a 1D periodic domain. Relative weightings are shown in the top plot, while the filter functions for this case are shown in the bottom plot. In the top plot, weightings used for $I_R^{\text{DI01}}$ and $I_R^{\text{D05}}$ are shown by the solid curves, while the dashed curves show the weightings used for $I_R^{\text{HL}}$ and $I_R^{\text{HL}}$. The observation errors are spatially uncorrelated and a SOAR model is used for the background-error correlations. The model parameters used are $L = 40,000 \text{ km}$, $L_b = 1,000 \text{ km}$, $\Delta x = 40 \text{ km}$, and $\sigma_b^2/\sigma_o^2 = 2$

since the square roots of the $s_b$ weightings are roughly inversely proportional to the innovation covariance filtered by $\tilde{F}$, modes at high wavenumbers are given much higher weights than at lower wavenumbers when determining $s_b$. The same is true for the $s_o$ weightings at lower wavenumbers, where the filter $I - \tilde{F}$ has low amplitudes and the weightings are large compared to those at the higher wavenumbers. This is evident for both the DI01 and D05 algorithms.

### 12 | CONCLUSIONS

We have shown that for the case of estimating error variances, the DI01, D05, and HL methods can all be
interpreted geometrically. The success of these methods are also dependent on the distribution of observations. For all methods, the innovation covariance must be sampled at an adequate resolution to resolve the background and observation correlations for the algorithm to be successful.

The analytic expressions for the expectation values and variances for the scaling parameter estimates from each method were determined for the ideal case in which the prescribed covariances differ from their true value solely by a scaling factor. For the DI01 and D05 methods, these expressions indicate that the divergence of the expectation values of the scaling parameters from their true values is proportional to quantities that measure the amount of spectral overlap between the background- and observation-error covariances. In this case, when the observation errors are uncorrelated, the expectation values of the HL scaling parameters equal their true values, but diverge from their true values when observation-error correlations are present. An examination of the expectation values for this case also shows that, if one scaling parameters is overestimated (either the background- or the observation-error variance scaling), then the other scaling parameter will be underestimated.

The statistical properties of the variance scaling parameter estimates were illustrated using the example of homogeneous covariances defined on a 1D periodic domain for the case when the correlations are properly specified. For this example, results showed that the expectation value of the background scaling parameter estimated using the D05 method was closer to its true value than that estimated using the DI01 method, although the D05 estimate had a larger variance than the DI01 method. On the other hand, the expectation value of the observation-error variance scaling parameter for the DI01 was closer to its true value than that using the D05 method, with the variance of both estimates being similar. For all cases examined, the HL scaling parameters had greater variance of the parameter estimates than both the DI01 and D05 methods.

Although the examples examined in this work were idealized cases, they can provide guidance for real-world applications. By using the actual spatial distribution in two or three dimensions for an observational dataset and estimates of the correlation length-scales of the background and observation errors, the derived analytical expressions for the expectation values and variances of the error variance scaling parameters can be estimated. These estimates can be used to compare different properties of each method for that particular observational dataset. For instance, one could determine which method yields the best estimates for the expectation values or the lowest variances of the scaling estimates. One could also use these expressions to check if the preferred method is dependent on which error variance is being tuned; for example, in the examples used in Section 10, it was found that the expectation value of the D05 background-error scaling parameter was closer to its true value than the DI01 estimate, but the opposite was true for the observation-error scaling parameter. Therefore, in applications where only one parameter (either the background or observation scaling) is focused on, the expressions in Section 10 and Appendix D can be used to suggest which method will be more beneficial for that particular situation.

The conceptual framework developed in this work (the results of Sections 5 to 7) does not assume any particular form (homogeneous, isotropic, or otherwise) for the true or modelled error covariances, other than for the HL method which assumes uncorrelated observation errors. A particular form for the true covariances was only introduced in Section 9 so that the performance of each method could be evaluated through a comparison of the tuned error variances to their true values. This was accomplished by assuming that the modelled error covariances were equal to their true values, expect for the case of the HL estimates for the example illustrated in Figure 10 which had observation errors with non-negligible spatial correlations. We note that, in cases where the error correlations are not properly specified, there is some ambiguity when defining the true error variance scaling factors which were defined in Section 9 and used to evaluate the performance of each method. For instance, one may define the true error variance scalings as the parameters that best fit the modelled error variances to the true error variances. On the other hand, one could alternatively define the true error variance scalings as the parameters which minimize the difference between the full modelled and true error covariances; in other words, parameters $s_b^2$ and $s_o^2$ that minimize $\|B^* - \tilde{s}_b B\|^2$ and $\|R^* - \tilde{s}_o R\|^2$ with some appropriate weighting. As such, the decomposition of the error covariance mis-specification into variance and correlation length-scaling parameters is not unique and care must be taken when interpreting results where both true variance and correlation parameters are defined.

As previously noted, although this work concentrates on refining estimates of the error variances, certain aspects of this work can be generalized for parametrizations of the error correlations. For instance, the generalized DI01 relations as expressed in Equations (14a) and (14b) and the D05 cost functions as written in Equations (21a) and (21b) are independent of parametrization and so could be used for correlation estimations. However, unlike for scalings of the error variances, in these cases the cost functions in Equations (21a) and (21b) will in general not be quadratic in terms of the parameter variables and the best-fit parameters emitted from their minimization may not have simple analytical expressions. While the absence of simple analytical expressions may make the fitting processes more...
difficult to interpret, nonetheless in these cases best-fit correlation parameters may be obtained by numerically minimizing the cost functions in Equations (21a) and (21b). As with the error variance scalings, the particular weights chosen for the cost functions will affect both the best-fit correlation parameter values obtained and the ease of implementing the algorithm in practical settings.

ACKNOWLEDGEMENTS
The authors would like to thank Yves Rochon and Martin Deshaies-Jacques for providing feedback on this paper.

ORCID
Michael Sitwell © https://orcid.org/0000-0003-4518-6278

REFERENCES
Bannister, R.N. (2008) A review of forecast-error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast-error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134, 1951–1970

Bathmann, K. (2018) Justification for estimating observation-error covariances with the Desroziers diagnostic. *Quarterly Journal of the Royal Meteorological Society*, 144, 1965–1974

Bennett, A.F. (1992) *Inverse Methods in Physical Oceanography*. Cambridge University Press, Cambridge, UK.

Bennitt, G.V., Johnson, H.R., Weston, P.P., Jones, J. and Potiaux, E. (2017) An assessment of ground-based GNSS zenith total delay observation errors and their correlations using the Met Office UKV model. *Quarterly Journal of the Royal Meteorological Society*, 143, 2436–2447

Bormann, N. and Bauer, P. (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: methods and application to ATOVS data. *Quarterly Journal of the Royal Meteorological Society*, 136, 1036–1050

Bormann, N., Collard, A. and Bauer, P. (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: application to AIRS and IASI data. *Quarterly Journal of the Royal Meteorological Society*, 136(649), 1051–1063

Chapnik, B., Desroziers, G., Rabier, F. and Talagrand, O. (2004) Properties and first application of an error-statistics tuning method in variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 130, 2253–2275

Chapnik, B., Desroziers, G., Rabier, F. and Talagrand, O. (2006) Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Quarterly Journal of the Royal Meteorological Society*, 132, 543–565

Cordoba, M., Dance, S.L., Kelly, G.A., Nichols, N.K. and Waller, J.A. (2017) Diagnosing atmospheric motion vector observation errors for an operational high-resolution data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 143, 333–341

Daescu, D.N. and Langland, R.H. (2013) Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. *Quarterly Journal of the Royal Meteorological Society*, 139, 226–241

Daescu, D.N. and Todling, R. (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Quarterly Journal of the Royal Meteorological Society*, 136, 2000–2012

Daley, R. (1991) *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, UK.

Daley, R. (1992) The lagged innovation covariance: a performance diagnostic for atmospheric data assimilation. *Monthly Weather Review*, 120(1), 178–196

Dee, D.P. and Da Silva, A.M. (1999) Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: methodology. *Monthly Weather Review*, 127(8), 1822–1834

Dee, D.P., Gaspari, G., Redder, C., Rukhovets, L. and Da Silva, A.M. (1999) Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: applications. *Monthly Weather Review*, 127(8), 1835–1849

Desroziers, G. (2012). Observation error specification, Chapter 8 in Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics. Blayo, É., Bocquet, M., Cosme, E., Cugliandolo, L.F. (eds), Oxford University Press, Oxford, UK.

Desroziers, G. and Ivanov, S. (2001) Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127, 1433–1452

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. (2005) Diagnosis of observation, background- and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131, 3385–3396

Eresmaa, R. and Järvinen, H. (2005) Estimation of spatial global positioning system zenith delay observation error covariance. *Tellus A*, 57(2), 194–203

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

Feng, J., Toth, Z., Peña, M. and Zhang, J. (2020) Partition of analysis- and forecast-error variance into growing and decaying components. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3738

Garand, L., Heilliette, S. and Buehner, M. (2007) Interchannel error correlation associated with AIRS radiance observations: inference and impact in data assimilation. *Journal of Applied Meteorology and Climatology*, 46(6), 714–725

Gauthier, P., Du, P., Heilliette, S. and Garand, L. (2018) Convergence issues in the estimation of interchannel correlated observation errors in infrared radiance data. *Monthly Weather Review*, 146(10), 3227–3239

Gupta, A.K. and Nagar, D.K. (2018) *Matrix Variate Distributions*. Chapman and Hall/CRC, Abingdon, UK.

Hollingsworth, A. (1986) Objective analysis for numerical weather prediction. Technical Memorandum 128, ECMWF, Reading, UK.

Hollingsworth, A. and Lonnberg, P. (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: the wind field. *Tellus A*, 38(2), 111–136

Horn, R.A. and Johnson, C.R. (1991) *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Hotta, D., Kalnay, E., Ota, Y. and Miyoshi, T. (2017) EFSR: ensemble forecast sensitivity to observation error covariance. *Monthly Weather Review*, 145(12), 5015–5031
Janjić, T., Bormann, N., Bocquet, M., Carton, J.A., Cohn, S.E., Dance, S.L., Losa, S.N., Nichols, N.K., Potthast, R., Waller, J.A. and Weston, P. (2018) On the representation error in data assimilation. Quarterly Journal of the Royal Meteorological Society, 144, 1257–1278

Julier, S.J. and Uhlmann, J.K. (2004) Unscented filtering and nonlinear estimation. Proceedings of the IEEE, 92(3), 401–422

Langland, R. and Baker, N. (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus A, 56(3), 189–201

Li, H., Kalnay, E. and Miyoshi, T. (2009) Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Quarterly Journal of the Royal Meteorological Society, 135, 523–533

Lorenc, A.C. (1986) Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112, 1177–1194

Macpherson, S. and Laroche, S. (2019) Estimation of ground-based GNSS zenith total delay temporal observation-error correlations using data from the NOAA and E-GVAP networks. Quarterly Journal of the Royal Meteorological Society, 145, 513–529

Ménard, R. (2016) Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks. Quarterly Journal of the Royal Meteorological Society, 142, 257–273

Ménard, R. and Chang, L.-P. (2000) Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: $\chi^2$-validated results and analysis of variance and correlation dynamics. Monthly Weather Review, 128(8), 2672–2686

Ménard, R., Deshaies-Jacques, M. and Gasset, N. (2016) A comparison of correlation-length estimation methods for the objective analysis of surface pollutants at Environment and Climate Change Canada. Journal of the Air and Waste Management Association, 66(9), 874–895. https://doi.org/10.1080/10962247.2016.1177620

Ménard, R., Gauthier, P., Rochon, Y., Robichaud, A., de Grandpré, J., Yang, Y., Charrette, C. and Chabrillat, S. (2019) Coupled stratospheric chemistry–meteorology data assimilation. Part II: weak and strong coupling. Atmosphere, 10(12), 798

Noble, B. (1969) Applied Linear Algebra. Pearson, Cambridge, UK.

Parrish, D.F. and Derber, J.C. (1992) The National Meteorological Center’s spectral statistical-interpolation analysis system. Monthly Weather Review, 120(8), 1747–1763

Peña, M. and Toth, Z. (2014) Estimation of analysis and forecast error variances. Tellus A, 66(1), 2176

Polavarapu, S., Ren, S., Rochon, Y., Sankey, D., Ek, N., Koshyk, J. and Tarasick, D. (2005) Data assimilation with the Canadian middle atmosphere model. Atmosphere–Ocean, 43(1), 77–100

Rutherford, I.D. (1972) Data assimilation by statistical interpolation of forecast error fields. Journal of the Atmospheric Sciences, 29(5), 809–815

Seber, G.A.F. and Lee, A.J. (2003) Linear Regression Analysis. John Wiley and Sons, Hoboken, NJ.

Skachko, S., Ménard, R., Errera, Q., Christophe, Y. and Chabrillat, S. (2016) ENKF and 4D-Var data assimilation with chemical transport model BASCOE (version 05.06). Geoscientific Model Development, 9(8), 2893–2908. https://doi.org/10.5194/gmd-9-2893-2016

Stewart, L.M., Dance, S.L., Nichols, N.K., Eyre, J.R. and Cameron, J. (2014) Estimating interchannel observation-error correlations for IASI radiance data in the Met Office system. Quarterly Journal of the Royal Meteorological Society, 140, 1236–1244

Talagrand, O. (1999). A posteriori evaluation and verification of analysis and assimilation algorithms, in Proceedings of Workshop on Diagnosis of Data Assimilation Systems, 2–4 November 1998. ECMWF, Reading, UK.

Talagrand, O. (2010). Evaluation of assimilation algorithms, pp. 217–240 in Data Assimilation: Making Sense of Observations. Lahoz, W.A., Khattatov, B., Ménard, R. (eds), Springer, Berlin.

Tarantola, A. (1987) Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation. Elsevier, Amsterdam, Netherlands.

Tongling, R. (2015) A complementary note to ‘A lag-1 smoother approach to system-error estimation’: the intrinsic limitations of residual diagnostics. Quarterly Journal of the Royal Meteorological Society, 141, 2917–2922

Waller, J.A., Dance, S.L. and Nichols, N.K. (2016a) Theoretical insight into diagnosing observation-error correlations using observation-minus-background and observation-minus-analysis statistics. Quarterly Journal of the Royal Meteorological Society, 142, 418–431

Waller, J.A., Simonin, D., Dance, S.L., Nichols, N.K. and Ballard, S.P. (2016b) Diagnosing observation-error correlations for Doppler radar radial winds in the Met Office UKV model using observation-minus-background and observation-minus-analysis statistics. Monthly Weather Review, 144(10), 3533–3551

Weston, P.P., Bell, W. and Eyre, J.R. (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. Quarterly Journal of the Royal Meteorological Society, 140, 2420–2429

---

**How to cite this article:** Sitwell M, Ménard R. Framework for the comparison of a priori and a posteriori error variance estimation and tuning schemes. *Q J R Meteorol Soc.* 2020;146:2547–2575. https://doi.org/10.1002/qj.3805

**APPENDIX A: SIMULTANEOUS DIAGONALIZATION OF COVARIANCE MATRICES**

In this appendix, we describe the process for constructing the operator which simultaneously diagonalizes both $\text{HBH}^T$ and $\hat{R}$. This standard mathematical procedure is described in mathematics textbooks such as Noble (1969) and in the context of data assimilation can be found in Hollingsworth (1986).
The operator $S$ is constructed in two steps: First, $\mathbf{HBH}^T$ is decomposed into its eigendecomposition as

$$\mathbf{HBH}^T = \mathbf{U}\Lambda_b\mathbf{U}^T,$$  \hspace{1cm} (A1)

which is also its singular value decomposition since $\mathbf{HBH}^T$ is a symmetric positive-definite matrix. In this decomposition, $\mathbf{U}$ is a unitary matrix which has the eigenvectors of $\mathbf{HBH}^T$ as its columns and $\Lambda_b$ is the corresponding diagonal matrix of eigenvalues. The operator $\mathbf{U}\Lambda_b^{-1/2}$ would then transform $\mathbf{HBH}^T$ into the identity matrix. Denote the transformation of $\mathbf{R}$ by $\mathbf{U}\Lambda_b^{-1/2}$ as $\mathbf{R}' \equiv \Lambda_b^{-1/2}\mathbf{U}^T\mathbf{R}\Lambda_b^{-1/2}$. The second part of the procedure is to find a unitary matrix $\mathbf{T}$ that diagonalizes $\mathbf{R}'$, which exists as $\mathbf{R}'$ is a covariance matrix. As $\mathbf{HBH}^T$ was transformed into the identity matrix by the operator $\mathbf{U}\Lambda_b^{-1/2}$, a subsequent transformation by $\mathbf{T}$ would have no effect on the transformation of $\mathbf{HBH}^T$. The total transformation is then defined as $S = \mathbf{U}\Lambda_b^{-1/2}\mathbf{T}$, which transforms $\mathbf{HBH}^T$ into the identity matrix and $\mathbf{R}$ into a diagonal matrix. Note that while $\mathbf{U}$ and $\mathbf{T}$ are unitary, $S$ is not.

**APPENDIX B: MATHEMATICAL IDENTITIES AND RELATIONS**

Mathematical identities and relations used in this paper are reviewed in this appendix. While most relations will be stated without proof, proofs can readily be found within mathematics textbooks (e.g., Horn and Johnson 1991).

- The Kronecker product $\mathbf{A} \otimes \mathbf{B}$ between the $m \times n$ matrix $\mathbf{A}$ and $p \times q$ matrix $\mathbf{B}$ is given by the $(mp) \times (nq)$ block matrix
  
  $$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,p}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m,1}\mathbf{B} & \cdots & a_{m,p}\mathbf{B} \end{bmatrix}, \hspace{1cm} (B1)$$

  where $a_{ij}$ denotes the element of $\mathbf{A}$ at the $i$th row and the $j$th column.

- For any matrices $\mathbf{A}$ and $\mathbf{B}$, the transpose over the Kronecker product is distributive, so that
  
  $$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T. \hspace{1cm} (B2)$$

- The product $\mathbf{A} \otimes \mathbf{B}$ is invertible if and only if $\mathbf{A}$ and $\mathbf{B}$ are both invertible, in which case the inverse is given by
  
  $$\mathbf{(A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}. \hspace{1cm} (B3)$$

- For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ of dimensions such that the products $\mathbf{AC}$ and $\mathbf{BD}$ can be formed, we have
  
  $$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}). \hspace{1cm} (B4)$$

- The vectorization operator $\text{vec} \ (\cdot)$ is a matrix operator that converts a matrix into a column vector by stacking the columns of the matrix on top of one another, so that the vectorization of the $m \times n$ matrix $\mathbf{A} = \begin{bmatrix} a_1 \ a_2 \ \ldots \ a_n \end{bmatrix}$ is given by
  
  $$\text{vec}(\mathbf{A}) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \hspace{1cm} (B5)$$

  where $a_1, a_2, \ldots, a_n$ are column vectors and $\text{vec} \ (\mathbf{A})$ has a length of $m \times n$.

- For any matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, the vectorization of the product $\mathbf{ABC}$ can be expressed as
  
  $$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B}). \hspace{1cm} (B6)$$

- Consider the trace $\text{Tr} \ [\mathbf{ABCD}]$ of the product of matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$, all with dimensions $n \times n$. From Equation (10), we have
  
  $$\text{Tr}[\mathbf{ABCD}] = \text{Tr} \ [(\mathbf{B}^T\mathbf{A}^T)^T\mathbf{CD}] = \text{vec}(\mathbf{B}^T\mathbf{A}^T)^T \text{vec}(\mathbf{CD}). \hspace{1cm} (B7)$$

  Using Equation (B6), we can write the trace as

  $$\text{Tr}[\mathbf{ABCD}] = \text{vec}(\mathbf{A}^T)^T \text{vec}(\mathbf{CD}) = \text{vec}(\mathbf{A}^T)^T (\mathbf{I}_n \otimes \mathbf{B})(\mathbf{D}^T \otimes \mathbf{I}_n)\text{vec}(\mathbf{C}), \hspace{1cm} (B8)$$

  where $\mathbf{I}_n$ is the identity matrix of dimension $n \times n$ and in the second line we have used the property in Equation (B2). Lastly, we use the property in Equation (B4) to simplify the trace to

  $$\text{Tr}[\mathbf{ABCD}] = \text{vec}(\mathbf{A}^T)^T (\mathbf{D}^T \otimes \mathbf{B})\text{vec}(\mathbf{C}). \hspace{1cm} (B9)$$

- The Hadamard product, also known as the Schur or element-wise product, between two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension is defined as

  $$\mathbf{(A \circ B)}_{ij} = a_{ij}b_{ij}, \hspace{1cm} (B10)$$

  where $a_{ij}$ and $b_{ij}$ denote the elements of $\mathbf{A}$ and $\mathbf{B}$, respectively. The product $\mathbf{A \circ B}$ has the same dimensions as $\mathbf{A}$ and $\mathbf{B}$.

- For a random (not necessarily normal) vector $\mathbf{x}$ with mean $\mu$ and covariance matrix $\Sigma$, the expectation value
of the quadratic form $x^T Ax$ with a symmetric matrix $A$ is

$$E[x^T Ax] = \text{Tr}[A\Sigma] + \mu^T A\mu. \quad (B11)$$

In the case that $A$ is not symmetric, then $x^T Ax = x^T A_x x$, where $A_x = (A + A^T)/2$ is the symmetric part of $A$, so for this case Equation (B11) holds if on the right-hand side $A$ is replaced by the symmetric part of $A$.

- For a random normal vector $x$ with mean $\mu$ and covariance matrix $\Sigma$, the variance of the quadratic form $x^T Ax$ with a symmetric matrix $A$ is

$$V[x^T Ax] = 2\text{Tr}[A\Sigma\Sigma] + 4\mu^T A\mu. \quad (B12)$$

As with Equation (B11), if $A$ is not symmetric, then Equation (B12) holds if on the right-hand side $A$ is replaced by the symmetric part of $A$.

**APPENDIX C: MINIMIZATION OF COST FUNCTIONS**

This section contains details relating to the minimization of quadratic cost functions of the forms encountered in the main text. Consider a cost function $J$ that is defined by

$$J = \frac{1}{2}s\|A - B\|^2_{W_1 \otimes W_2}$$

$$= \frac{1}{2}\text{Tr}[(sA - B)W_2 (sA - B)W_1]$$

$$= \frac{1}{2}s^2\text{Tr}[AW_1AW_2] - s\text{Tr}[AW_1BW_2]$$

$$+ \frac{1}{2}\text{Tr}[BW_1BW_2]. \quad (C1)$$

where $A$ and $B$ are symmetric matrices, $s$ is a scalar, and $W_1$ and $W_2$ are positive definite weightings. The second line in Equation (C1) follows from the definition of the weighted Frobenius inner product given in Equation (11) and in the third line we have used the cyclic property of the trace operator. The derivative of $J$ with respect to $s$ is given by

$$\frac{\partial J}{\partial s} = s \text{Tr}[AW_1AW_2] - \text{Tr}[AW_1BW_2] \quad (C2)$$

Setting this derivative equal to zero and solving for $s$ yields

$$\hat{s} = \frac{\text{Tr}[AW_1BW_2]}{\text{Tr}[AW_1AW_2]}\quad (C3)$$

The second derivative of the cost function $J$ with respect to $s$ is

$$\frac{\partial^2 J}{\partial s^2} = \text{Tr}[AW_1AW_2] = \|A\|^2_{W_1 \otimes W_2}. \quad (C4)$$

As this second derivative is positive for all values of $s$, $s = \hat{s}$ is the minimum of $J$.

**APPENDIX D: VARIANCES OF SCALING PARAMETER ESTIMATES**

This appendix contains expressions for the variances of scaling parameter estimates derived by the DI01, D05, and HL methods, introduced in Sections 5,6, and 7, respectively. These variances are calculated under the same assumptions as used for calculating the expectation values in Section 9, namely that the modelled covariances differed from their true values by only an overall scaling factor. The variances are calculated by use of Equation (B12) in Appendix B, which, unlike when calculating the expectation values, assume that the innovations $d$ follow a normal distribution.

For the DI01 algorithm, the variances of $s_b^{\text{DI01}} / s^o_b$ and $s_o^{\text{DI01}} / s^o_b$ are given by

$$V\left(\frac{s_b^{\text{DI01}}}{s_b^o}\right) = \frac{2\text{Tr}\left[(HB^T D^{-1}(HB^T + \frac{\xi}{\bar{\xi}} R) D^{-1})^2\right]}{\text{Tr}[HB^T D^{-1}]}$$

$$= \frac{2\text{Tr}\left[(I+\Phi)^{-4} + \frac{\xi}{\bar{\xi}} \Phi(I+\Phi)^{-4} + \left(\frac{\xi}{\bar{\xi}}\right)^2 \Phi^2(I+\Phi)^{-4}\right]}{\text{Tr}[\Phi(I+\Phi)^{-4}]}. \quad (D1a)$$

$$V\left(\frac{s_o^{\text{DI01}}}{s_b^o}\right) = \frac{2\text{Tr}\left[(RD^{-1}(\frac{\xi}{\bar{\xi}} HB^T + R) D^{-1})^2\right]}{\text{Tr}[RD^{-1}]}$$

$$= \frac{2\text{Tr}\left[\Phi^4(I+\Phi)^{-4} + \frac{\xi}{\bar{\xi}} \Phi(I+\Phi)^{-4} + \left(\frac{\xi}{\bar{\xi}}\right)^2 \Phi^2(I+\Phi)^{-4}\right]}{\text{Tr}[\Phi(I+\Phi)^{-4}]}. \quad (D1b)$$

where in the last lines we have used Equations (3a) and (3b) to transform into the simultaneously diagonalized basis. For the D05 algorithm, the variances are given by

$$V\left(\frac{s_b^{\text{D05}}}{s_b^o}\right) = \frac{\text{Tr}\left[(HB^T D^{-1} + D^{-1}HB^T)(HB^T + \frac{\xi}{\bar{\xi}} R)\right]}{2\text{Tr}[HB^T]}. \quad (D2a)$$
\[ V \left( \frac{s_{05}^2}{s_0^2} \right) = \frac{\text{Tr} \left\{ \left( \begin{array}{cc} 0 & -D^{-1} R \\ D^{-1} & 1 \end{array} \right) \left( \begin{array}{cc} 0 & H BH^T + R \\ H BH^T & 0 \end{array} \right) \right\}}{2\text{Tr}[R]} \]  
\quad \text{(D2b)}

For the HL method, the variances are given by
\[ V \left( \frac{s_{\text{HL}}^2}{s_b^2} \right) = \frac{2\text{Tr} \left\{ \left( \begin{array}{cc} \Sigma^{-1} & 0 \\ 0 & \Sigma^{-1} \end{array} \right) \left( \begin{array}{cc} V & 0 \\ 0 & (V \circ D) \Sigma^{-1} \end{array} \right) \left( \begin{array}{cc} 0 & B \\ B & 0 \end{array} \right) \right\}}{\text{Tr} \left[ (V \circ D) \Sigma^{-2} (V \circ D) \Sigma^{-2} \right]} \]  
\quad \text{(D3a)}

where
\[ \Omega = \frac{\text{Tr} \left[ (I \circ R) \Sigma^{-2} (I \circ B) \Sigma^{-2} \right]}{\text{Tr} \left[ (V \circ D) \Sigma^{-2} (V \circ D) \Sigma^{-2} \right]} \]  
\quad \text{(D3b)}