University of Reading

# *Treating sample covariances for use in strongly coupled atmosphere-ocean data assimilation*

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Smith, P. J. ORCID: https://orcid.org/0000-0003-4570-4127, Lawless, A. S. ORCID: https://orcid.org/0000-0002-3016-6568 and Nichols, N. K. ORCID: https://orcid.org/0000-0003-1133-5220 (2018) Treating sample covariances for use in strongly coupled atmosphere-ocean data assimilation. Geophysical Research Letters, 45 (1). pp. 445-454. ISSN 0094-8276 doi: 10.1002/2017gl075534 Available at https://centaur.reading.ac.uk/74247/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1002/2017gl075534

Publisher: American Geophysical Union

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2017GL075534

**Key Points:**
- Methods are compared for improving the rank of multivariate sample forecast error covariance matrices for coupled data assimilation systems
- Reconditioning retains correlation structure but does not remove sample noise; localization removes sample noise but can destroy structure
- A third strategy that combines the advantages of the reconditioning and state-space localization approaches leads to improved results

**Supporting Information:**
- Supporting Information S1
- Supporting Information S2
- Supporting Information S3
- Supporting Information S4
- Data Set S1

**Correspondence to:**
P. J. Smith,
p.j.smith@reading.ac.uk

**Citation:**
Smith, P. J., Lawless, A. S., & Nichols, N. K. (2018). Treating sample covariances for use in strongly coupled atmosphere-ocean data assimilation. *Geophysical Research Letters*, *45*, 445–454. https://doi.org/10.1002/2017GL075534

Received 2 SEP 2017
Accepted 5 DEC 2017
Accepted article online 12 DEC 2017
Published online 15 JAN 2018

©2017. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

# Treating Sample Covariances for Use in Strongly Coupled Atmosphere-Ocean Data Assimilation

**Polly J. Smith**[1], **Amos S. Lawless**[1,2], **and Nancy K. Nichols**[1,2]

[1]School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, UK, [2]National Centre for Earth Observation, University of Reading, Reading, UK

**Abstract** Strongly coupled data assimilation requires cross-domain forecast error covariances; information from ensembles can be used, but limited sampling means that ensemble derived error covariances are routinely rank deficient and/or ill-conditioned and marred by noise. Thus, they require modification before they can be incorporated into a standard assimilation framework. Here we compare methods for improving the rank and conditioning of multivariate sample error covariance matrices for coupled atmosphere-ocean data assimilation. The first method, reconditioning, alters the matrix eigenvalues directly; this preserves the correlation structures but does not remove sampling noise. We show that it is better to recondition the correlation matrix rather than the covariance matrix as this prevents small but dynamically important modes from being lost. The second method, model state-space localization via the Schur product, effectively removes sample noise but can dampen small cross-correlation signals. A combination that exploits the merits of each is found to offer an effective alternative.

## 1. Introduction

The ideal approach to coupled data assimilation is strongly (or fully) coupled in which components of the Earth system are analyzed within a single seamless assimilation framework: a single assimilation algorithm is used to combine the available observations from each subcomponent with a fully coupled model forecast state, producing a single fully coupled analysis. Despite recent advancements (e.g., Frolov et al., 2016; Laloyaux et al., 2016; Lea et al., 2015; Sluka et al., 2016) there are still a number of scientific challenges to solve before coupled assimilation can become the primary technique for weather and climate forecasting and reanalysis (Penny & Hamill, 2017). A particular challenge for strongly coupled data assimilation systems is modeling the cross covariances that characterize the relationship between errors in the atmosphere and ocean model forecasts (Smith et al., 2017).

Our initial work (Smith et al., 2015) with an idealized single column coupled atmosphere-ocean model and incremental 4D-Var assimilation framework found that even when the errors in the initial atmosphere and ocean forecast states are assumed to be uncorrelated (i.e., cross-domain error covariances set to zero) the strongly coupled approach generally outperforms both the standard uncoupled and intermediate weakly coupled formulations in terms of producing more balanced initial analysis fields and reducing initialization shock in subsequent forecasts. This is attributed to the ability of the strongly coupled 4D-Var algorithm to implicitly generate atmosphere-ocean error cross covariances (see, e.g., Bannister, 2008) thereby enabling observations in one fluid to directly influence the analysis increments in the other. It is expected that explicitly prescribing nonzero atmosphere-ocean cross-domain forecast error covariances a priori will have further positive impact on the assimilation in that it will allow near-surface observations to be used to even greater effect and thus lead to even greater consistency between the atmosphere and ocean analysis states.

The problem of 4D-Var data assimilation is to find the initial state such that the model forecast best fits the available observations over a given time window, while remaining close to a given a priori guess (the "background") and allowing for the uncertainty in each. This best estimate (the "analysis") should be consistent with both the observations and the system dynamics. Estimates of the statistics of the errors in the background state and observations are described by the background (or forecast) and observation error covariance matrices, **B** and **R**, respectively. Variational methods have traditionally used a static or climatological estimate of **B**, but more recently, methods have been developed to incorporate flow-dependent ensemble

information into variational frameworks (see, e.g., the review by Bannister, 2017, and references therein). A covariance matrix derived from a sample of size $m$ can have at most rank $(m-1)$ and so can only be full rank if $m > n$ where $n$ is the dimension of the state space. Given that the size of most ensembles is restricted by computational resources, ensemble error covariances will routinely be rank deficient; further, when the size of the ensemble is significantly smaller than the dimension of the state the effect of sampling errors is likely to be nonnegligible.

For uncoupled ensemble-based data assimilation systems, various localization techniques have been designed to treat the problems associated with small ensembles, in both model and observation space (see, e.g., Ménétrier et al., 2015, for a review of current methods). Standard 4D-Var formulations work in model space and require $\mathbf{B}$ to be symmetric positive definite (so all eigenvalues are strictly positive) so that $\mathbf{B}^{-1}$ exists. By construction, a sample covariance/correlation matrix will be symmetric positive semidefinite; this means that it may have zero eigenvalues and hence be numerically rank deficient. Even if a matrix is positive definite, it may still be poorly conditioned and hence not accurately invertible (Golub & Van Loan, 2013). In either case, before we can use such a matrix in a standard data assimilation framework it will need to be modified.

The purpose of this letter is to address these issues in the context of strongly coupled atmosphere-ocean data assimilation; specifically, we investigate two strategies for treating fully coupled sample error covariance matrices that are rank deficient and/or ill-conditioned and evaluate their effectiveness in terms of preserving the underlying error correlation structures, particularly those between the atmosphere and ocean. The first approach, which we term "reconditioning," is to perform an eigenvector decomposition and alter the spectrum of the matrix directly: making any zero eigenvalues positive and moving the largest and smallest eigenvalues closer together; the matrix is then reformed using the modified eigenvalues and original eigenvectors (see, e.g., Daniels & Kass, 2001; Ledoit & Wolf, 2004 for discussion of variations of this type of approach, and Weston et al., 2014 and Campbell et al., 2017 for recent examples of its application). The second option is to use localization techniques (e.g., Houtekamer & Mitchell, 2001). A standard approach to model space localization is to use the Schur product (element-wise multiplication) of the raw ensemble forecast error covariance matrix and a positive-definite correlation matrix, $\boldsymbol{\rho} \in \mathbb{R}^{n \times n}$; this process acts to remove the zero eigenvalues of the raw ensemble forecast error covariance/correlation matrix and thus increase its rank. Although localization in model space is the natural choice for variational-based data assimilation, the size of the coupled model state, the contrast in scales between the atmosphere and ocean, plus the use of different coordinate systems all combine to make the coupled localization problem more complex. Both the strategies we present can be used to ensure that the matrix is strictly positive definite and at the same time improve its conditioning; improving the conditioning will in turn improve the convergence and stability of the 4D-Var minimization.

## 2. Matrix Modification Methods

### 2.1. Reconditioning

There are various ways in which the eigenvalues of a matrix can be modified. The simplest approach is to set a minimum threshold, $\lambda_{\text{tol}}$, and reset all eigenvalues less than this value equal to it (e.g., Weston et al., 2014). Since only the smallest eigenvalues are changed, this method modifies the overall eigenvalue structure. As discussed in Weston et al. (2014), resetting the smallest eigenvalues to the same fixed value results in many of the small diagonal values of the original matrix being reset to an almost constant value. An alternative approach is to specify a required condition number, $\kappa_{\text{tol}}$, and increment all eigenvalues by a fixed amount $\lambda_{\text{inc}}$ such that

$$\frac{\lambda_{\text{max}} + \lambda_{\text{inc}}}{\lambda_{\text{min}} + \lambda_{\text{inc}}} = \kappa_{\text{tol}}\,, \tag{1}$$

where $\lambda_{\text{max}}$, $\lambda_{\text{min}}$ are the largest and smallest eigenvalues, respectively. Although incrementing all eigenvalues means that the largest eigenvalues are changed, the effect is relatively modest compared to the effect on the smallest eigenvalues; furthermore, this method has the advantage that the overall structure of the eigenvalues is preserved.

Note that reconditioning a covariance matrix as opposed to a correlation matrix will have differing effects on the underlying correlation structures. As we discuss in section 4.1, this result is particularly important for coupled systems due to the range of scales involved.

### 2.2. Localization

Localization is used as a way of filtering out spurious correlations from ensemble covariances when small ensembles are used but can also be designed to increase the effective ensemble size and ensure that the covariance matrix is full rank (Hamill et al., 2001; Oke et al., 2007); therefore, it can potentially also be used as a way of improving the conditioning of ensemble error covariance/correlation matrices that are derived from a large ensemble size (and so relatively noise free) but are rank deficient or poorly conditioned. Typically, the localization matrix $\boldsymbol{\rho}$ is assumed to be an isotropic distance-dependent function such that the amplitude of the error correlation between two locations decreases as the distance between them increases and falls to zero beyond a given separation; the rate of correlation decrease and cutoff depends on a user defined length-scale parameter.

The development of a coupled atmosphere-ocean vertical localization function presents two key issues: first, the atmosphere and ocean model components may use different vertical coordinate systems and have very different localization length scales; second, the majority of preexisting localization methods have been developed for uncoupled systems and are focused on a single state variable, or use control variable transform/balance operator techniques that convert to a new set of uncorrelated variables (e.g., Kepert, 2009). The cross media atmosphere-ocean vertical localization problem is naturally multivariate, and the assumption that the errors in the atmosphere and ocean model forecasts are correlated is a fundamental one.

A popular form for $\boldsymbol{\rho}$ in meteorological applications is the 5th-order piecewise rational function described by Gaspari and Cohn (1999) (their equation (4.10), hereafter referred to as GC). As it is written, this is a univariate localization function and it is easy to see how the entries of $\boldsymbol{\rho}$ might be constructed for a single state variable on a discrete 1-D model grid; it is less immediately obvious how to construct $\boldsymbol{\rho}$ when the model state vector is multivariate. Roh et al. (2015) use a simple bivariate model to investigate two strategies for the design of localization functions for multivariate state vectors. The first method is based on forming a multivariate localization matrix from the univariate GC function. A similar approach is adopted by Frolov et al. (2016) for localizing the coupled cross-domain ensemble covariances in their atmosphere-ocean interface solver. One potential flaw in this approach, noted by Roh et al. (2015), is that a multivariate localization matrix constructed this way is not guaranteed to be positive definite; they propose fixing this by setting $\boldsymbol{\rho}$ equal to the product of a univariate function and a symmetric, positive-definite matrix whose diagonal entries are one. However, this acts to further attenuate cross-variable correlations and so is undesirable for cases where these correlations are small but significant.

For coupled atmosphere-ocean assimilation, the importance of intervariable error correlations is well acknowledged, especially in the lower atmosphere-upper ocean boundary layer; but these correlations may be small relative to single variable autocorrelations and to correlations between errors in variables within the same fluid; therefore, extra care must be taken to that ensure that the localization does not destroy the smaller, yet physical, atmosphere-ocean error cross correlations. We have developed a vertical localization strategy built on similar ideas to those proposed by Roh et al. (2015) and Frolov et al. (2016). Our approach is the same on a broad level, but here we give specific consideration to the trade-off between filtering sampling noise and improving conditioning, and preserving the atmosphere-ocean error cross-correlation signals; this aspect of coupled model space localization is not addressed in previous studies.

Unlike the reconditioning approach, localization via the Schur product can be applied to either the raw error covariance matrix or error correlation matrix with the same effect; for ease of presentation we focus our discussion on the ensemble forecast error correlations. We partition the ensemble atmosphere-ocean forecast error correlation matrix into blocks

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathbf{AA}} & \mathbf{C}_{\mathbf{AO}} \\ \mathbf{C}_{\mathbf{AO}}^{T} & \mathbf{C}_{\mathbf{OO}} \end{pmatrix} \in \mathbb{R}^{n \times n}, \tag{2}$$

where $n = n_a + n_o$ is the dimension of the combined atmosphere-ocean model state vector. The blocks $\mathbf{C}_{\mathbf{AA}} \in \mathbb{R}^{n_a \times n_a}$ and $\mathbf{C}_{\mathbf{OO}} \in \mathbb{R}^{n_o \times n_o}$ are square matrices representing the atmosphere and ocean state forecast error correlations, respectively. The off-diagonal block $\mathbf{C}_{\mathbf{AO}} \in \mathbb{R}^{n_a \times n_o}$ contains the cross correlations between errors in the atmosphere and ocean state variables. These blocks can be further decomposed into submatrices containing the error autocorrelations for individual variables and error cross correlations between different variables. Note that neither $\mathbf{C}_{\mathbf{AO}}$ nor its submatrices are square. We apply a separate GC-type localization function to each submatrix so our multivariate $\boldsymbol{\rho}$ has the same block structure as (2).

## 3. The Coupled 1-D Model System

To compare the methods described in sections 2.1 and 2.2, we use sample coupled error covariances and correlations that were generated using an idealized strongly coupled single-column atmosphere-ocean incremental 4D-Var assimilation framework. In this section we provide a brief overview of this system; a full description of the model and strongly coupled incremental 4D-Var assimilation algorithm is given in Smith et al. (2015); details of the ensemble methodology used to estimate the coupled forecast error correlations are presented in Smith et al. (2017).

### 3.1. The Model

The model consists of a simplified version of the European Centre for Medium-range Weather Forecasts atmosphere single column model, which is based on an early cycle of their Integrated Forecast System code, coupled to a 1-D mixed layer ocean model developed by the National Centre for Atmospheric Science Centre for Global Atmospheric Modelling (Woolnough et al., 2007) and based on the K-Profile Parametrization (KPP) vertical mixing scheme of Large et al. (1994). The two components communicate via the exchange of sea surface temperature and surface fluxes of heat, moisture, and momentum. The full model equations are presented in Smith et al. (2015), together with the model validation. A detailed discussion of the different interactions between the atmosphere and ocean model variables and their errors is presented in Smith et al. (2017).

### 3.2. Strongly Coupled Incremental 4D-Var Data Assimilation

Solving the full nonlinear 4D-Var problem directly can be a complex and expensive procedure; the incremental formulation (Courtier et al., 1994) circumvents these issues by linearizing about the current state estimate and replacing the nonlinear problem with a sequence of linear least squares problems. Rather than searching for the optimal initial state directly, it searches for increments to the initial background estimate; this is done iteratively via a series of linearized inner-loop quadratic cost function minimizations and nonlinear outer-loop update steps (see, e.g., Lawless et al., 2005). For strongly coupled incremental 4D-Var the control vector consists of both the atmosphere and ocean prognostic variables, the observation vector contains both atmosphere and ocean observations, and the coupled model is used in both the inner and outer loops.

### 3.3. Ensemble Derived Forecast Error Correlations

In a follow up (Smith et al., 2017) to our initial study (Smith et al., 2015), estimates of the atmosphere-ocean error cross correlations for our system were derived using the analysis ensemble methodology of Zagar et al. (2005). This approach generates samples of background error by taking short-range forecast differences between members of a data assimilation ensemble; the statistics of these differences are then used as a proxy for the forecast error covariance matrix. Experiments were identical twin; each ensemble consisted of 500 cycled strongly coupled incremental 4D-Var data assimilations, with each member starting from a different randomly perturbed initial background state and assimilating a different set of randomly perturbed observations over eight consecutive 12 h cycles. An ensemble of 500 was chosen after looking at the convergence of the correlation structures for successively larger ensemble sizes. For the purposes of this paper, we consider the effect of modifying the raw ensemble error covariance/correlation matrix derived for a winter case in which the true state is given by a coupled model forecast from 00 UTC on 2 December 2013; the matrix was computed by averaging 12 h forecast differences between pairs of analysis ensemble members from cycles 2, 4, 6, and 8. Both the covariance and correlation matrices have zero eigenvalues and so are rank deficient; they also clearly contain sampling noise.

## 4. Results

### 4.1. Matrix Reconditioning

For our system the smallest eigenvalues are mainly associated with the ocean where variability is low compared to the atmosphere. Increasing the smallest eigenvalues (either to the same constant value or incrementally) mainly has the effect of weakening the error correlations in and between the ocean fields, and the error cross correlations between the ocean and atmosphere; the larger the value of $\lambda_{\text{tol}}$, or the smaller the specified $\kappa_{\text{tol}}$, the bigger the increase in the smallest eigenvalues and the greater the effect on the ocean error correlations. This effect is particularly pronounced when we recondition the error covariance matrix as opposed to the error correlation matrix, as illustrated in Figure 1 for the case $\kappa_{\text{tol}} = 10^4$. The differences between the original and reconditioned correlation matrix are very small (the minimum and maximum

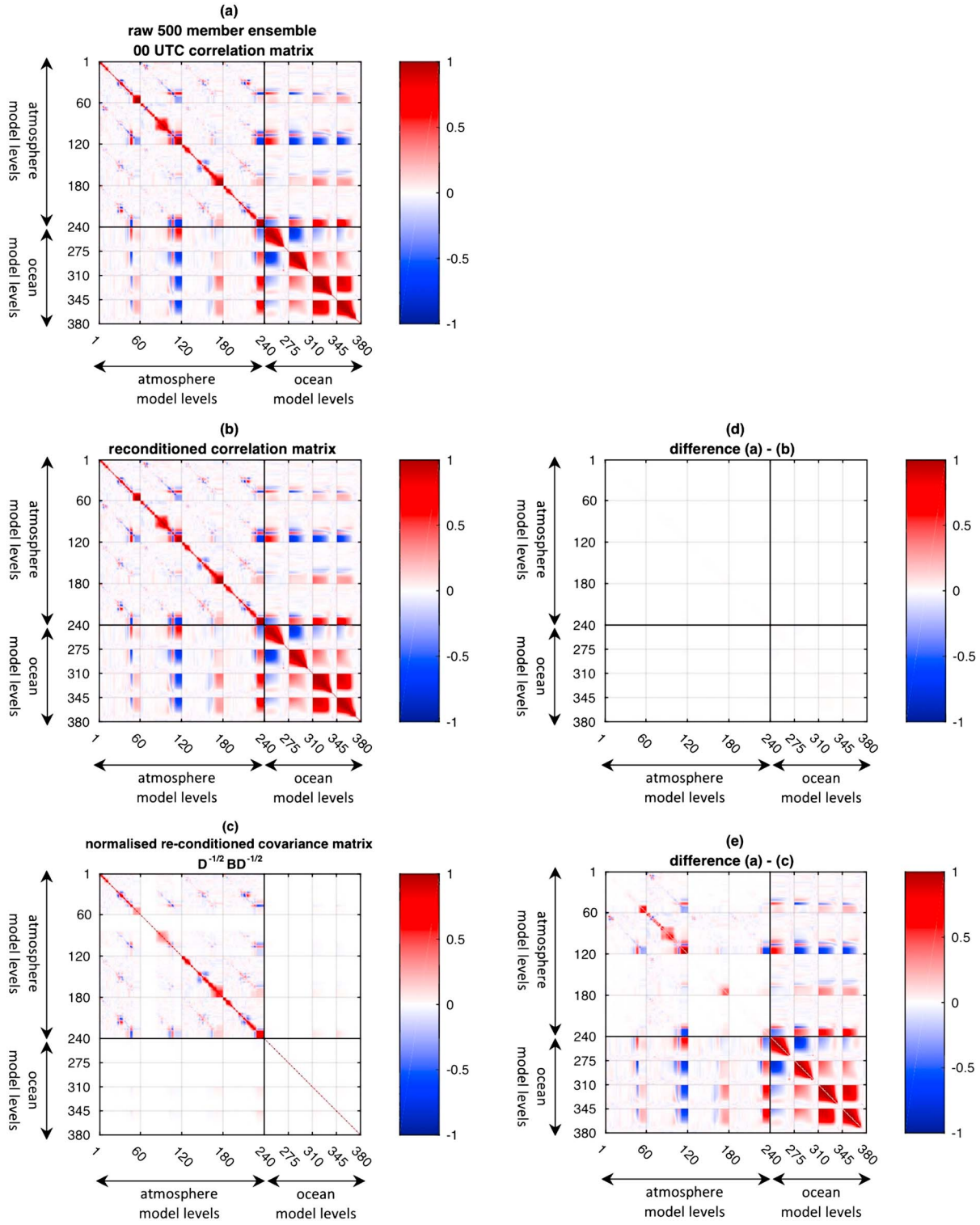


**Figure 1.** Comparison of the structure of (a) the original ensemble error correlation matrix, (b) the reconditioned error correlation matrix, and (c) the normalized reconditioned error covariance matrix when $\kappa_{tol} = 10^4$. (d and e) Show the difference between Figures 1a and 1b, 1c. Model levels 1–240 correspond to atmosphere temperature, specific humidity, and zonal and meridional wind on 60 levels; model levels 241–380 correspond to ocean temperature, salinity, and zonal and meridional current on 35 levels.

differences in Figure 1d are $\sim \mathcal{O}(10^{-3})$); the structure of the correlations has been retained but so too has the sampling noise. Figure 1c shows the result of reconditioning the covariance matrix and then normalizing by premultiplying and postmultiplying with a diagonal matrix of inverse forecast error standard deviations. Here the structure of the ocean cross correlations and the atmosphere-ocean cross-correlation structure has been almost completely lost.

The largest and smallest (nonzero) eigenvalues of the ensemble error covariance and correlation matrices will not be the same and will not necessarily correspond to the same modes of variability, that is, the directions of greatest variance are not necessarily the same as the directions of greatest correlation, and so modifying the eigenvalues of the error covariance matrix versus the error correlation matrix will have differing effects on the correlation structures. In order to retain as much of the atmosphere-ocean cross-domain error correlation information as possible, it is better to modify the ensemble correlation matrix rather than the covariance matrix. This highlights an important point: even though a particular eigenvector may be associated with a small eigenvalue of the covariance matrix (and therefore assumed to account for little variance) it may still be dynamically important, particularly for coupled systems where the individual components have very different space and time scales. Care should therefore be taken with methods that truncate matrix eigenvalues, such as reduced order assimilation techniques.

### 4.2. Localization

Construction of the localization matrix $\boldsymbol{\rho}$ requires specification of a distance measure and decorrelation length scales. Generally, the localization distance is measured as the Euclidean distance between model grid points. Our ocean model uses a fixed vertical grid that stretches from 1 to 250 m depth, and so computing the distance between two ocean model grid points (for subblocks of $\boldsymbol{\rho}_{\boldsymbol{OO}}$) is simple. The atmosphere model uses a 60 level hybrid coordinate system that depends on the surface pressure; however, this is prescribed and does not vary significantly over the period of our experiments. We therefore assign each model level a fixed mean pressure value which we convert to a set of altitudes; this enables us to define the distance between atmosphere model levels in terms of difference in height (m) and makes the definition of the distance between the atmosphere and ocean points (for subblocks of $\boldsymbol{\rho}_{\boldsymbol{AO}}$) straightforward. If we denote the height at level $i$ ($i = 1, \ldots, 60$) of the atmosphere model as $z_a(i)$, and the depth of level $j$ ($j = 1, \ldots, 35$) of the ocean model as $z_o(j)$, element $\rho(i,j)$ of each submatrix of $\boldsymbol{\rho}_{\boldsymbol{AO}}$ will then depend on the distance

$$d(z_a(i), z_o(j)) = z_a(i) + z_o(j)\,, \tag{3}$$

since atmosphere heights are positive upward and ocean depths are positive downward.

For the localization length scales, we should, ideally, specify a separate value for each individual state variable and for each combination, or pair, of variables. However, initial experiments showed that the positive definiteness of the constructed $\boldsymbol{\rho}$ matrix depended on how these values were chosen. In general, using different length scales for different subblocks of $\boldsymbol{\rho}$ failed to produce even a semidefinite matrix whereas using a single fixed value for all atmosphere variables in $\boldsymbol{\rho}_{\boldsymbol{AA}}$, and similarly for all ocean variables in $\boldsymbol{\rho}_{\boldsymbol{OO}}$ did. Although we found some variations in the ensemble error correlation length scales from one variable to another within each fluid, they were all of the same order of magnitude and so this approach is not unreasonable. Further, we found that if $\boldsymbol{\rho}$ and $\mathbf{C}$ are both positive semidefinite then their Schur product can be strictly positive definite; however, we are not currently aware of any theorems that state the conditions for which this result is guaranteed.

For the block $\mathbf{C}_{\mathbf{AO}}$ we are primarily interested in retaining the atmosphere-ocean error cross correlations within the near-surface atmosphere-ocean boundary layer, as this is where they are strongest (Smith et al., 2017). For our model, the atmosphere ensemble error correlation length scales are approximately 2 orders of magnitude greater than those for the ocean over the time scales we consider; to ensure that the atmospheric variables do not have undue influence on the ocean beyond the near-surface boundary, we need to account for this disparity when we define the localization length scales for the subblock $\boldsymbol{\rho}_{\boldsymbol{AO}}$. Similarly to Frolov et al. (2016), we found that this could be achieved by replacing (3) with the scaled distance

$$\hat{d}(z_a(i), z_o(j)) = \left( \frac{z_a(i)}{L_a} + \frac{z_o(j)}{L_o} \right)\,, \tag{4}$$

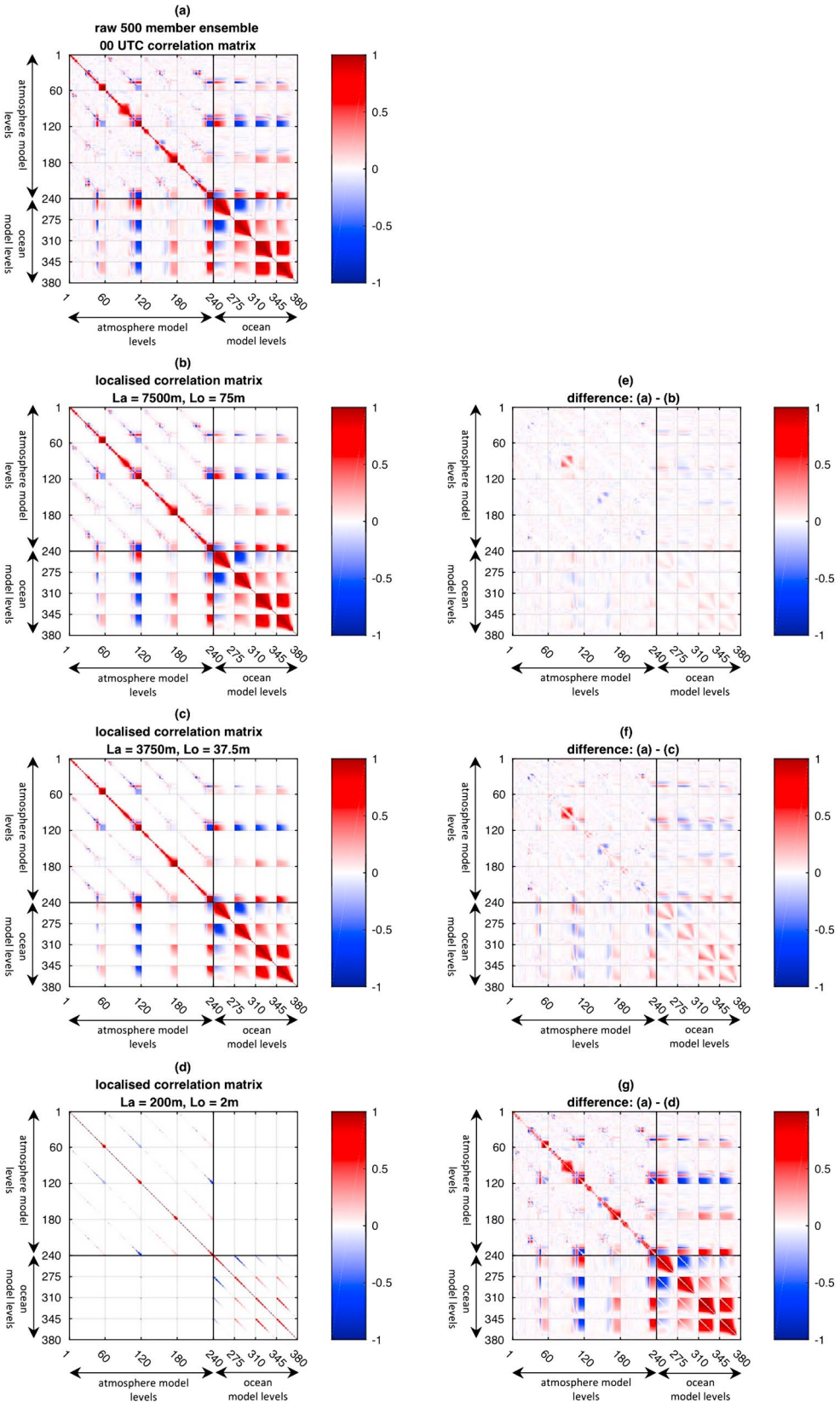


**Figure 2.** Comparison of the structure of (a) the original ensemble error correlation matrix and the localized error correlation matrix for length scales: (b) $L_a = 7,500$ m, $L_o = 75$ m; (c) $L_a = 3,750$ m, $L_o = 37.5$ m; and (d) $L_a = 200$ m, $L_o = 2$ m. (e–g) Show the difference between Figures 2a and 2b–2d. Model levels are as in Figure 1.

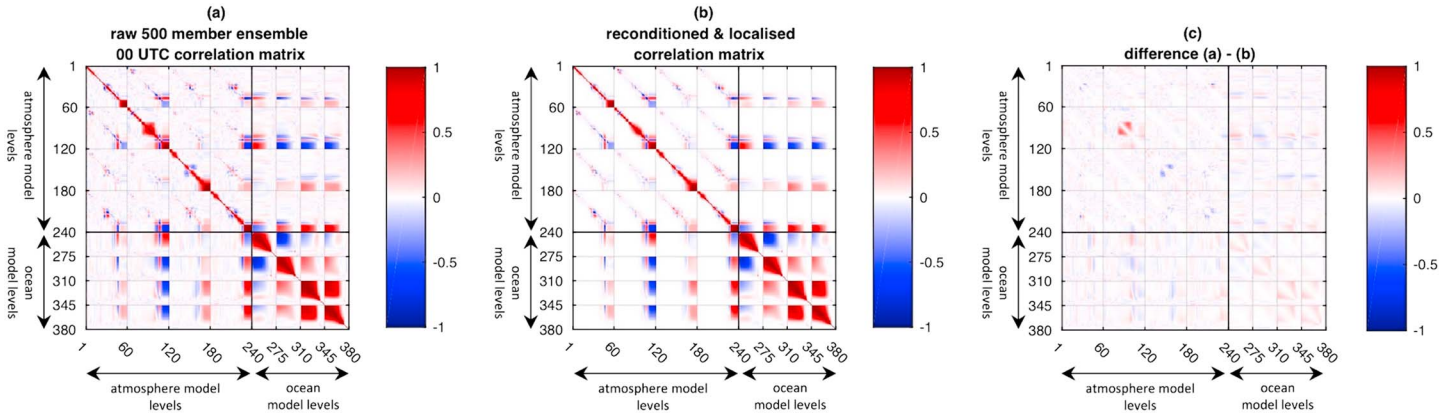


**Figure 3.** Comparison of the structure of (a) the original ensemble error correlation matrix and (b) error correlation matrix that has been reconditioned by setting $\kappa_{\text{tol}} = 10^4$ and then localized using $L_a = 7,500$ m and $L_o = 75$ m. (c) The difference between Figures 3a and 3b. Model levels are as in Figure 1.

where $L_a$ and $L_o$ represent the height and depth of the near-surface atmosphere-ocean boundary region, respectively. The elements of $\boldsymbol{\rho}_{AO}$ then take the form (cf. GC equation (4.10))

$$\rho_{AO}(\hat{d}) = \begin{cases} -\frac{1}{4}\hat{d}^5 + \frac{1}{2}\hat{d}^4 + \frac{5}{8}\hat{d}^3 - \frac{5}{3}\hat{d}^2 + 1\,, & 0 \le \hat{d} \le 1 \\ \frac{1}{12}\hat{d}^5 - \frac{1}{2}\hat{d}^4 + \frac{5}{8}\hat{d}^3 + \frac{5}{3}\hat{d}^2 - 5\hat{d} + 4 - \frac{2}{3}\hat{d}^{-1}\,, & 1 \le \hat{d} \le 2 \\ 0\,, & 2 \le \hat{d}\,. \end{cases} \tag{5}$$

When both $z_a > L_a$ and $z_o > L_o$, $\rho_{AO} = 0$; when $z_a < L_a$and $z_o < L_o$, $\rho_{AO}$ will be close to 1 for points near the surface and decrease as $z_a \to L_a$ and/or $z_o \to L_o$. If $L_a < z_a < 2L_a$but $z_o < L_o$then $\rho_{AO}$ will be nonzero but small and become smaller as $z_o \to L_o$ or $z_a \to 2L_a$; a similar argument holds when $z_a < L_a$and $L_o < z_o < 2L_o$. Appropriate values for $L_a$ and $L_o$ will depend on a number of factors including the ensemble size, model system, and forecast period. In this case, choosing short correlation length scales reduces the condition number of the localized matrix but overdamps the correlation structures, particularly those in the atmosphere-ocean surface region, meaning that smaller correlations are lost. Unlike the reconditioning approach we cannot choose the condition number or minimum eigenvalue. Here localization produces a bigger adjustment to the larger eigenvalues in addition to increasing the smallest eigenvalues and so has a much greater impact on the overall eigenvalue structure; it is not possible to reduce the condition number to the same extent using this approach without completely destroying the matrix cross correlations. As an example, Figure 2 compares the raw and localized error correlation matrices for the cases $L_a = 7,500$ m, $L_o = 75$ m (Figures 2b and 2e) and $L_a = 3,750$ m, $L_o = 37.5$ m (Figures 2c and 2f). Halving the length scales reduces the condition number, $\kappa(\mathbf{C})$, by a factor of 10 from $\mathcal{O}(10^9)$ to $\mathcal{O}(10^8)$ but the strength of the cross correlations are noticeably reduced. We found that we could obtain $\kappa(\mathbf{C}) \sim \mathcal{O}(10^4)$ by reducing the length scales to $L_a = 200$ m, $L_o = 2$ m but the localized correlations are quite obviously incorrect (Figures 2d and 2g).

### 4.3. Combining Reconditioning and Localization

Another potential strategy is to employ a combination of the first two: first reconditioning the raw error correlation matrix to produce a required condition number and then localizing to remove noise; since the reconditioned matrix will already be positive definite this will allow more flexibility in the choice of localization function and length scales. Figure 3 shows the result of this approach; first reconditioning using $\kappa_{\text{tol}} = 10^4$ and then localizing using $L_a = 7,500$ m, $L_o = 75$ m. Comparing this with the results of the individual methods (Figures 1 and 2) shows that this combined approach produces a matrix that is visually similar to the localized correlation matrix in Figure 2b in that the sampling noise has been filtered (compare Figures 2e and 3c). However, the matrix in Figure 3b has condition number $\mathcal{O}(10^4)$, as opposed to $\mathcal{O}(10^9)$ for the matrix in Figure 2b; the reconditioning step has enabled the matrix condition number to be reduced without destroying the correlation structure.

## 5. Summary

In this study we have compared methods for improving the rank and conditioning of multivariate sample forecast error covariance matrices in the context of strongly coupled atmosphere-ocean data assimilation:

reconditioning whereby the matrix eigenvalues are altered directly, and model space localization via the Schur product. There are advantages and disadvantages to each of these strategies. Assuming the computational effort required to perform the eigen-decomposition is not restrictive, matrix reconditioning is simple: it does not require changes to the model coordinate system or tuning of length scales and so offers a practical option when sampling noise is not an issue, for example, when the sample size is large relative to the dimension of the state. However, as illustrated, modifying the eigenvalues of an error covariance matrix as opposed to the associated error correlation matrix will have differing effects on the underlying correlation structures. Care needs to be taken with the choice of $\kappa_{tol}$ or $\lambda_{tol}$ particularly when applying the approach to systems with a range of scales. An important finding from this study is that methods that retain only the leading eigenpairs after decomposition, such as empirical orthogonal functions (EOFs), may not be optimal for representing the dynamics of a coupled system. At the very least, consideration needs to be given to the scaling of the problem to ensure that eigenvalues that are small but correspond to dynamically significant modes are not truncated.

Within the data assimilation community, localization is often the default choice for treating noisy, rank deficient ensemble covariances. However, the standard model space Schur product technique does not immediately translate to coupled systems; here we have presented a potential approach in the context of our 1-D system. Compared to reconditioning, Schur product localization is very effective at removing spurious noise, but because it is a more blanket approach it can also destroy true error correlation signals that are small. Because a pair of atmosphere and ocean points in a layered model will always have vertical separation, the scaled distance $\hat{d}$ in (5) can never be zero and $\rho_{AO}(\hat{d})$ will always be less than 1; therefore, any cross-domain error correlations will always be damped by this type of distance-dependent localization. This can be remedied to a certain extent by an appropriate choice of atmosphere and ocean length scales, but the GC based localization function presented here is restrictive in its current form in that it does not allow us to account for differences in the correlation length scales for different components of the state vector. There are strategies proposed in the statistics literature for constructing positive-definite localization matrices with inhomogeneous length scales that could potentially be used to overcome this limitation (see, e.g., Kleiber & Porcu, 2015; Paciorek & Schervish, 2006). There is also a trade-off between the length scales and the conditioning of the localized matrix; shorter length scales give a lower condition number but also wipe out the off-diagonal correlation matrix elements and so are useless in this context.

Finally, we presented a third strategy that combines the benefits of the first two approaches; using reconditioning to reduce the matrix condition number and then localizing to reduce the sampling noise. Obviously, this approach will be computationally more expensive, but it offers greater flexibility in that it places less restriction on the choice of correlation length scales when a certain level of conditioning is desired. This may prove to be useful in the development of multivariate localization functions with multiple length scales.

Naturally, the topic of cross-fluid localization requires further exploration and will no doubt ultimately demand a far more sophisticated approach for full 3-D systems; nonetheless, this study has highlighted some crucial issues that will be important to build upon going forward.

**Acknowledgments**
This work is funded by the Natural Environment Research Council (NERC) grant NE/M001482/1. The data and code used to produce the results in this article are included as supporting information.

## References

Bannister, R. N. (2008). A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Quarterly Journal of Royal Meteorological Society*, *134*, 1971–1996. https://doi.org/10.1002/qj.340

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 607–633. https://doi.org/10.1002/qj.2982

Campbell, W., Satterfield, E., Ruston, B., & Baker, N. (2017). Accounting for correlated observation error in a dual formulation 4D-Variational data assimilation system. *Monthly Weather Review*, *145*(3), 1019–1032. https://doi.org/10.1175/MWR-D-16-0240.1

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

Daniels, M. J., & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, *57*(4), 1173–1184. https://doi.org/10.1111/j.0006-341X.2001.01173.x

Frolov, S., Bishop, C., Holt, T., Cummings, J., & Kuhl, D. (2016). Facilitating strongly-coupled ocean-atmosphere data assimilation with an interface solver. *Monthly Weather Review*, *144*(1), 3–20. https://doi.org/10.1175/MWR-D-15-0041.1

Gaspari, G., & Cohn, S. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*, 723–757. https://doi.org/10.1002/qj.49712555417

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*. Baltimore: Johns Hopkins University Press.

Hamill, T., Whitaker, J., & Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, *129*(11), 2776–2790. https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2

Houtekamer, P., & Mitchell, H. (2001). A sequential Ensemble Kalman Filter for atmospheric data assimilation. *Monthly Weather Review*, *129*, 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2

Kepert, J. D. (2009). Covariance localisation and balance in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, *135*(642), 1157–1176. https://doi.org/10.1002/qj.443

Kleiber, W., & Porcu, E. (2015). Nonstationary matrix covariances: Compact support, long range dependence and quasi-arithmetic constructions. *Stochastic Environmental Research and Risk Assessment*, *29*(1), 193–204. https://doi.org/10.1007/s00477-014-0867-6

Laloyaux, P., Balmaseda, M., Dee, D., Mogensen, K., & Janssen, P. (2016). A coupled data assimilation system for climate reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *142*, 65–78. https://doi.org/10.1002/qj.2629

Large, W., McWilliams, J., & Doney, S. (1994). Oceanic vertical mixing: A review and a model with non-local boundary layer parameterization. *Reviews of Geophysics*, *32*, 363–403. https://doi.org/10.1029/94RG01872

Lawless, A. S., Gratton, S., & Nichols, N. K. (2005). An investigation of incremental 4D-Var using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society*, *131*, 459–476. https://doi.org/10.1256/qj.04.20

Lea, D. J., Mirouze, I., Martin, M. J., King, R. R., Hines, A., Walters, D., & Thurlow, M. (2015). Assessing a new coupled data assimilation system based on the Met Office coupled atmosphere-land-ocean-sea ice model. *Monthly Weather Review*, *143*(11), 4678–4694. https://doi.org/10.1175/MWR-D-15-0174.1

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Ménétrier, B., Montmerle, T., Michel, Y., & Berre, L. (2015). Linear filtering of sample covariances for ensemble-based data assimilation. Part I: Optimality criteria and application to variance filtering and covariance localization. *Monthly Weather Review*, *143*(5), 1622–1643. https://doi.org/10.1175/MWR-D-14-00157.1

Oke, P., Sakov, P., & Corney, S. (2007). Impacts of localisation in the EnKF and EnOI: Experiments with a small model. *Ocean Dynamics*, *57*(1), 32–45. https://doi.org/10.1007/s10236-006-0088-8

Paciorek, C. J., & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, *17*(5), 483–506. https://doi.org/10.1002/env.785

Penny, S. G., & Hamill, T. M. (2017). Coupled data assimilation for integrated earth system analysis and prediction. *Bulletin of the American Meteorological Society*, *98*(7), ES169–ES172. https://doi.org/10.1175/BAMS-D-17-0036.1

Roh, S., Jun, M., Szunyogh, I., & Genton, M. (2015). Multivariate localization methods for ensemble Kalman filtering. *Nonlinear Processes in Geophysics Dicussions*, *2*, 833–863. https://doi.org/10.5194/npgd-2-833-2015

Sluka, T., Penny, S., Kalnay, E., & Miyoshi, T. (2016). Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation. *Geophysical Research Letters*, *43*, 752–759. https://doi.org/10.1002/2015GL067238

Smith, P. J., Fowler, A. M., & Lawless, A. S. (2015). Exploring strategies for coupled 4D-Var data assimilation using an idealised atmosphere-ocean model. *Tellus A*, *67*(27025). https://doi.org/10.3402/tellusa.v67.27025

Smith, P. J., Lawless, A. S., & Nichols, N. K. (2017). Estimating forecast error covariances for strongly coupled atmosphere-ocean 4D-Var data assimilation. *Monthly Weather Review*, *145*(10), 4011–4035. https://doi.org/10.1175/MWR-D-16-0284.1

Weston, P. P., Bell, W., & Eyre, J. R. (2014). Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society*, *140*(685), 2420–2429. https://doi.org/10.1002/qj.2306

Woolnough, S., Vitart, F., & Balmaseda, M. (2007). The role of the ocean in the Madden-Julian Oscillation: Implications for MJO prediction. *Quarterly Journal of the Royal Meteorological Society*, *133*, 117–128. https://doi.org/10.1002/qj.4

Zagar, N., Andersson, E., & Fisher, M. (2005). Balanced tropical data assimilation based on a study of equatorial waves in ECMWF short-range forecast errors. *Quarterly Journal of the Royal Meteorological Society*, *131*, 987–1101. https://doi.org/10.1256/qj.04.54