

Correlation-cutoff method for covariance localization  
in strongly coupled data assimilation

Takuma Yoshida

A scholarly paper in partial fulfillment of the requirements for the degree of  
Master of Science  
November 2018

Department of Atmospheric and Oceanic Science,  
University of Maryland, College Park

Advisors: Prof. Eugenia Kalnay  
Prof. Stephen G. Penny

## Abstract

The use of coupled models is now considered essential for Earth science and seasonal forecast. The coupled models have shown their ability to reproduce natural variability of various timescales and its response to human activities. Taking advantage of this advancement, data assimilation with those coupled models has been studied to obtain the best state estimate of the coupled system. This approach is called coupled data assimilation. The earlier stage of coupled data assimilation, called weakly coupled data assimilation (WCDA), where the coupling takes place by just coupling the forecasts, has been successfully implemented in operational atmosphere and ocean analyses.

The more sophisticated method is called strongly coupled data assimilation (SCDA). In SCDA, the cross-covariance of background error is utilized for correcting the background state of the coupled system consistently. Previous studies of SCDA with ensemble Kalman filters (EnKFs) show contradicting results: some study obtained better analysis than WCDA, whereas other study reported the analysis was degraded than WCDA. Since the relationship between WCDA and SCDA is equivalent to the relationship between analyses with and without localization (Section 2.1), strongly coupled EnKF is naturally more susceptible to the rank deficiency and spurious correlation problems. Therefore, localization of the analysis is the central problem for the successful implementation of strongly coupled EnKF.

In this study, we first show that the square of background error correlation between the observed and analyzed variables plays a crucial role in the reduction of analysis uncertainty. This result suggests the use of correlation-cutoff method, in which we cut-off the assimilation of observations guided by formerly estimated error statistics. Experiments

of the method with a nine-variable coupled model with three subsystems show overall superior analysis compared to WCDA and standard SCDA with enhanced robustness to the ensemble size.

To obtain the optimal covariance localization for more realistic applications, we also examine the background error correlation structure of the global atmosphere-ocean system using a coupled general circulation model (Fast Ocean Atmosphere Model; FOAM). A physically plausible correlation is found between some variables of the atmosphere and the ocean, suggesting that we may improve the analysis of those variables by implementing SCDA. The implementation of the correlation-cutoff method to the coupled general circulation model remains for the future work.

## Table of Content

Abstract.....	2
List of Tables .....	5
List of Figures.....	6
List of Abbreviations .....	8
1. Introduction.....	9
2. Correlation-cutoff method and experiments with a nine-variable coupled model .....	14
2.1 Theoretical analysis .....	14
2.2 Correlation-cutoff method .....	16
2.3 Experimental settings.....	17
2.4 Experimental results .....	19
3. Experiments with a coupled general circulation model.....	26
3.1 Experimental settings.....	26
3.2 Background error correlation of a coupled GCM.....	29
4. Summary and Future Work .....	34
References .....	36

## List of Tables

Table 1: Specification of the Fast Ocean Atmosphere Model (FOAM). See Chapter 2 of Jacob (1997) for detail.....	28
--	----

List of Figures

Figure 1: Schematics of non-coupled data assimilation (left) and weakly and strongly coupled data assimilation (right). See Introduction for a detailed description..... 11

Figure 2: (a) Mean of squared error correlation for each pair of variables (Eq. 5) obtained by an offline LETKF experiment with the nine-variable model of Peña and Kalnay (2004). (b) Mean of squared error correlation for all the 81 ordered pairs of variables, sorted in descending order. Note the drop in correlation after the first 45 highly correlated pairs of variables, which indicates that the *ENSO-coupling* localization pattern in Figure 4 is optimal. Figure from Yoshida and Kalnay (2018)..... 20

Figure 3: Same with Figure 2(a) but (a) estimated with an LETKF cycle assimilating observations of all nine variables instead of only  $ye, yt$ , and  $Y$  and (b) estimated with a 4-member WCDA cycle instead of a 10-member fully coupled SCDA cycle.. 21

Figure 4: Covariance localization patterns tested. Background covariance is considered only at the shaded pairs of variables. Except for the offline experiment of Figure 3(a), only the observations (columns) indicated by “+” signs are assimilated into model variables (rows). Figure from Yoshida and Kalnay (2018) ..... 22

Figure 5: Temporal mean analysis root-mean-square error (RMSE) of each experiment. Different shading shows the different covariance localization pattern (see the legend and Figure 4). In each panel, the errors in the extratropical atmosphere, the tropical atmosphere, and the ocean are separately shown. Horizontal lines indicate the observation errors of each component for comparison. Each panel is the result of (a) 4-member experiments, (b) 6-member experiments, and (c) 10-member experiments. Note that the filter diverges and cannot finish all the cycle correctly in the 4-member

Full experiment. Figure from Yoshida and Kalnay (2018).....	24
Figure 6: Ensemble correlation within vertical columns averaged for 260 instances (00UTC of 260 days starting on model date January 1 <sup>st</sup> ). The panels show (a) ensemble correlation between meridional wind and zonal current at point 40°S 80°E (southern Indian Ocean) and (b) ensemble correlation between atmospheric and oceanic temperature at 0° 120°W (Eastern Equatorial Pacific). Deep ocean layers below the model topography are whited-out. Upper-left and lower-right quadrants of each panel show atmospheric and oceanic internal error correlation, respectively, and the other two quadrants show cross-correlation between the atmosphere and the ocean.....	31
Figure 7: Mean ensemble background correlation of (a) ocean surface zonal current to an ocean surface zonal current observation at point 40°S 80°E (southern Indian Ocean) and (b) atmospheric surface meridional wind to an atmospheric surface meridional wind observation at the same point, averaged for 260 instances (00UTC of 260 days starting on model date January 1 <sup>st</sup> ).....	33

## List of Abbreviations

DA	Data Assimilation
EnKF	Ensemble Kalman Filter
ENSO	El Niño Southern Oscillation
FOAM	Fast Ocean Atmosphere Model
GCM	General Circulation Model
IAU	Incremental Analysis Update
LACC	Leading Averaged Coupled Covariance
LETKF	Local Ensemble Transform Kalman Filter
MPI	Message Passing Interface
OM3	Ocean Model version 3
OSSE	Observation System Simulation Experiment
PCCM3	Parallel Community Climate Model version 3
RMSE	Root Mean Squared Error
SCDA	Strongly Coupled Data Assimilation
SST	Sea Surface Temperature
UTC	Coordinated Universal Time
WCDA	Weakly Coupled Data Assimilation

## 1. Introduction

In traditional numerical weather prediction, sea surface temperature (SST), land surface, sea ice, and chemical variables are treated as prescribed boundary conditions given to the atmospheric models. This approximation has been successful, as long as the target of the prediction is limited to the synoptic weather up to several days ahead. However, when we try to predict phenomena with longer timescales, the “boundary conditions” vary in time due to their internal processes and interactions with the other components. Furthermore, some natural variability modes, such as El Niño Southern Oscillation (ENSO), are intrinsically coupled modes which cannot be explained either by atmospheric or oceanic internal processes. To predict such coupled phenomena, coupled models that explicitly predict the atmosphere, the ocean, the cryosphere, and other biogeochemical processes are employed. They have shown an ability to predict the natural variability and its response to human activities.

Similarly, data assimilation (DA) has been developed to analyze each subsystem of the Earth individually; the atmospheric state is estimated using an atmospheric model and atmospheric observations, and so is the state of the ocean and other subsystems of the Earth. Such individually estimated analyses, when combined and used as initial conditions of a coupled model, may cause “initialization shock” and deteriorate the forecast due to dynamical inconsistencies (e.g., Chen et al., 1995; Zhang et al., 2007; Mulholland et al., 2015). Therefore, to provide dynamically consistent initial conditions to coupled models, *coupled data assimilation*, data assimilation with coupled models, has been studied (e.g., Penny and Hamill, 2017, and references therein). In addition to better initialization of coupled models, coupled DA is also expected to provide better constraints on subsystems in

which few measurements are available (Zhang et al., 2007). For example, global observations of the internal ocean were not been available before the deployment of Tao array (McPhaden et al., 1998) and ARGO floats (Argo, 2000), which is limiting our understanding of oceanic climate variability (Zhang et al., 2010).

Coupled DA is broadly divided into two stages of sophistication: weakly coupled data assimilation (WCDA) and strongly coupled data assimilation (SCDA) (Penny et al., 2017). Figure 1 shows schematics of conventional non-coupled and coupled atmosphere-ocean DA. In WCDA, a coupled model is used for updating short-term forecast or the background (blue boxes in Figure 1 right) whereas in non-coupled DA, an atmospheric/oceanic model uses pre-determined boundary conditions typically derive from an antecedent reanalysis or observations interpolated without dynamical models (dashed black arrows in Figure 1 left). Due to the employment of physical laws of coupled dynamics in the form of a coupled model, we expect that the background of WCDA is more self-consistent and accurate than the one produced by non-coupled models, and the improved background estimate will also improve the resulting analysis. Due to the coupling of forecast steps, analysis increments in a component can propagate to the other component during the forecast steps to indirectly correct the background of the next analysis.

In SCDA, in addition, we make use of the cross-covariance between the atmospheric and oceanic background errors (hereafter the term cross-covariance (cross-correlation) is used to refer to that of the background errors). This enables the error in the atmospheric and oceanic backgrounds to be corrected all at once by assimilation of observations. Consider, for example, we observed surface air temperature and knew that the background surface air temperature predicted by the coupled model was too warm compared to the observation. In

that case, we would imagine that the background SST would also be too warm compared to the truth. Strongly coupled DA enables to correct these errors in a self-consistent manner even if SST observation is unavailable (red arrows in Figure 1).

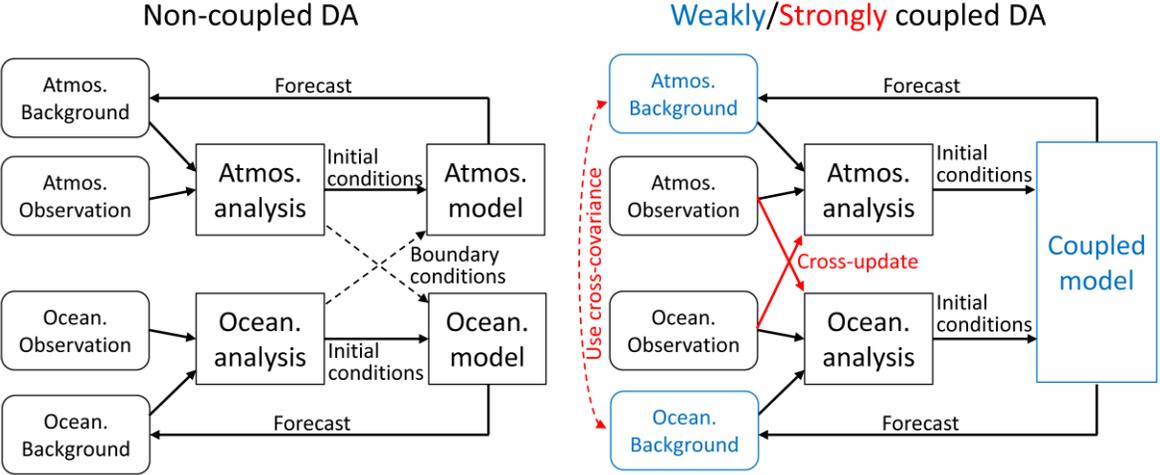


Figure 1: Schematics of non-coupled data assimilation (left) and weakly and strongly coupled data assimilation (right). See Introduction for a detailed description.

The earlier stage of coupled DA, WCDA, was successfully implemented as coupled real-time analyses and reanalyses of the atmosphere and the ocean (e.g., Saha et al., 2010; Laloyaux et al., 2016; Karspeck et al., 2018). Although the most sophisticated method, SCDA, is expected to provide even better analysis than WCDA, its effectiveness is still under investigation (Penny et al., 2017). Some of the important previous studies are listed below.

Sluka et al. (2016) assimilated simulated atmospheric observations directly into unobserved oceanic variables and showed that the analysis error in the ocean was reduced by 46% compared to WCDA using the local ensemble transform Kalman filter (LETKF;

Hunt et al., 2007) and a coupled general circulation model (GCM). They also conducted experiments to assimilate both atmospheric and oceanic observations, obtaining better analysis accuracy near the surface (Sluka, 2017, personal comm.). Han et al. (2013) tested strongly coupled ensemble Kalman filter (EnKF; Evensen, 1994) using a six-variable model that mimics the atmosphere, the ocean, and the sea ice. They showed that the use of cross-covariance between subsystems only improved the oceanic analysis accuracy if and only if they used a very large ensemble ( $\sim 10^4$ ); the atmospheric analysis could not be improved by directly assimilating oceanic observations. Kang et al. (2011) studied SCDA between dynamic and carbon dioxide variables. They showed that the best analysis was obtained when they ignored the cross-covariance between carbon dioxide flux/concentration and some dynamical variables like temperature or specific humidity. Their successful approach of ignoring some of the cross-covariances, called “variable localization”, is based on the observation that temperature and humidity, unlike wind, do not have a direct physical interaction with carbon dioxide flux/concentration.

Lu et al. (2015a, b) tested direct assimilation of atmospheric observations into oceanic variables and showed that simple SCDA is only beneficial in the deep tropics, where the ocean drives the atmosphere through anomalous sea surface temperature (Ruiz-Barradas et al., 2017). They attributed the detrimental effect of SCDA in higher latitudes to the fluctuating synoptic weather, which does not last long enough to strongly interact with the ocean. Therefore, they proposed the Leading Averaged Coupled Covariance (LACC) method, in which the innovation of atmospheric observations is averaged for several days to suppress the effect of the weather noise.

These existing studies of SCDA are contradictory, and this raises an important question: *Under what condition the direct assimilation of an observation into the other components (i.e., SCDA) improves the analysis compared to WCDA?* In other words, *what determines the optimal coupling/localization in the analysis of coupled systems?* We will answer this fundamental question by a theoretical/experimental study and explore the optimal assimilation strategy for SCDA.

## 2. Correlation-cutoff method and experiments with a nine-variable coupled model

### 2.1 Theoretical analysis

From a mathematical viewpoint, SCDA allows cross-covariances between background errors of components and their evolution through coupled forward models (e.g., Smith et al., 2017). In an example of a two-component system of the atmosphere and the ocean, the coupled background error covariance matrix  $\mathbf{B}_{\text{SCDA}}$  is written as

$$\mathbf{B}_{\text{SCDA}} = \begin{bmatrix} \mathbf{B}_{\text{AA}} & \mathbf{B}_{\text{AO}} \\ \mathbf{B}_{\text{AO}}^T & \mathbf{B}_{\text{OO}} \end{bmatrix}, \quad (1)$$

where the subscripts A and O denote the atmosphere and the ocean;  $\mathbf{B}_{\text{AA}}$ ,  $\mathbf{B}_{\text{OO}}$ , and  $\mathbf{B}_{\text{AO}}$  are the background error covariances within the atmosphere, within the ocean, and between the atmosphere and the ocean. In contrast, in the analysis steps of WCDA, each component is analyzed by itself, which means it solves the minimization problem using the background error covariance

$$\mathbf{B}_{\text{WCDA}} = \begin{bmatrix} \mathbf{B}_{\text{AA}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{\text{OO}} \end{bmatrix}, \quad (2)$$

ignoring the cross-covariances, or the off-diagonal blocks of  $\mathbf{B}_{\text{SCDA}}$  (Frolov et al., 2016). Therefore, the advancement from WCDA to SCDA is in some sense equivalent to removing the covariance localization, a necessary treatment for EnKFs with limited ensemble size (e.g., Hamill et al., 2001; Ott et al., 2004; Greybush et al., 2011). Hence it is understandable that a naïve implementation of a strongly coupled EnKF requires larger ensemble size than a weakly coupled EnKF as experimentally shown by Han et al. (2013).

However, in practice, we want to improve the analysis without increasing the ensemble size and the computational cost. To meet this requirement, one needs to couple only the “relevant” pairs of components but ignore unimportant or unreliable covariances as Kang et al. (2011) empirically did. This discussion motivates us to estimate how much analysis uncertainty of a model variable is reduced by assimilating each observation.

Under an assumption that each observation is assimilated sequentially with the Kalman filter (Kalman, 1960), the relative decrease of analysis uncertainty of each model variable can be written in a simple formula (Yoshida and Kalnay, 2018):

$$\frac{\sigma_{bi}^2 - \sigma_{ai}^2}{\sigma_{bi}^2} = \frac{\sigma_{yb}^2}{\sigma_{yb}^2 + \sigma_{yo}^2} \text{corr}^2(\delta x_{bi}, \delta y_b), \quad (3)$$

where  $\sigma_{bi}$  ( $\sigma_{ai}$ ) is the background (analysis) error standard deviation of the analyzed variable (indexed by  $i$ ),  $\sigma_{yb}$  ( $\sigma_{yo}$ ) is the background (observation) error standard deviation of the observed variable, and  $\text{corr}(\delta x_{bi}, \delta y_b)$  is the correlation between the background errors of the analyzed variable ( $\delta x_{bi}$ ) and the observed variable ( $\delta y_b$ ). The assumption of sequential assimilation of observation is not very restricting because the Kalman filter can assimilate each observation sequentially without changing the analysis if the observation errors are not correlated (e.g., Houtekamer and Mitchell, 2001), although the sequential assimilation of observations may change the analysis when localization is applied (Nerger, 2015; Kotsuki et al., 2017).

This simplified equation indicates that the improvement of the analysis at each model variable by a single observation is the product of two quantities: *(i) the ratio of the background and total error variances at the observed variable* (this ratio is close to one

when the observation is precise relative to the background) and (ii) *the square of background error correlation between the analyzed and observed variables*. This equation also provides a quantitative estimate of analysis error reduction by considering cross-covariances (i.e., the difference between SCDA and WCDA).

We hypothesize that in EnKFs, the assimilation of “irrelevant” observations deteriorates the coupled analysis when the detrimental effect of spurious correlation exceeds this expected error reduction in the Kalman filter. This argument suggests the use of a correlation-cutoff method in strongly coupled EnKFs, in which we only consider cross-covariance between variables that have strong background error correlation and cut-off the assimilation of irrelevant observations.

## 2.2 Correlation-cutoff method

Using an offline analysis cycle of a coupled EnKF, we first calculate error statistics as follows.

Assuming a constant observation network, for each pair of a model variable and an observable  $(x_i, y_j)$ , we first calculate an instantaneous background ensemble correlation at each analysis time  $t$ :

$$\text{corr}_{ij}(t) = \frac{\sum_{k=1}^K [x_{bi}^k(t) - \bar{x}_{bi}(t)][y_{bj}^k(t) - \bar{y}_{bj}(t)]}{\sqrt{\sum_{k=1}^K [x_{bi}^k(t) - \bar{x}_{bi}(t)]^2} \sqrt{\sum_{k=1}^K [y_{bj}^k(t) - \bar{y}_{bj}(t)]^2}}, \quad (4)$$

where  $x_{bi}^k$  and  $\bar{x}_{bi}$  respectively are background values of the  $i$ th model variable of the  $k$ th ensemble member and the ensemble mean (similarly  $y_{bj}^k$  and  $\bar{y}_{bj}$  are defined for the

$j$ th observation), and  $K$  is the ensemble size. Then for each pair  $(x_i, y_j)$ , we obtain the temporal mean of squared correlation:

$$\langle \text{corr}_{ij}^2 \rangle = \frac{1}{T} \sum_{t=1}^T \text{corr}_{ij}^2(t), \quad (5)$$

where  $T$  is the number of assimilation windows used to estimate the error statistics. If some of the model variables are directly observed, the observation  $y_j$  is just a model variable  $x_j$ , which is the case for our experiments in subsections 2.4 and 3.2. For a nonconstant observation network, the statistics can be aggregated for similar sets of observations (e.g., those with similar observation type, level, and latitude) instead of aggregating for each observation index  $j$ .

As suggested by Eq. (3), we can use the static quantities  $\langle \text{corr}_{ij}^2 \rangle$  to know the priority of assimilation of each observation into each model variable and cut-off the assimilation when the squared correlation becomes small.

### 2.3 Experimental settings

We have tested the correlation-cutoff method with the nine-variable toy coupled model of Peña and Kalnay (2004) with model equations:

$$\begin{aligned} \dot{x}_e &= \sigma(y_e - x_e) - c_e(Sx_t + k_1) \\ \dot{y}_e &= rx_e - y_e - x_e z_e + c_e(Sy_t + k_1) \\ \dot{z}_e &= x_e y_e - bz_e \\ \dot{x}_t &= \sigma(y_t - x_t) - c(SX + k_2) - c_e(Sx_e + k_1) \\ \dot{y}_t &= rx_t - y_t - x_t z_t + c(SY + k_2) + c_e(Sy_e + k_1) \end{aligned} \quad (6)$$

$$\begin{aligned}
\dot{z}_t &= x_t y_t - b z_t + c_z Z \\
\dot{X} &= \tau \sigma (Y - X) - c(x_t + k_2) \\
\dot{Y} &= \tau r X - \tau Y - \tau S X Z + c(y_t + k_2) \\
\dot{Z} &= \tau S X Y - \tau b Z - c_z z_t,
\end{aligned}$$

where the parameters are  $(\sigma, r, b, \tau, c, c_z, c_e, S, k_1, k_2) = (10, 28, 8/3, 0.1, 1, 1, 0.08, 1, 10, -11)$ , and the integration time step is  $\Delta t = 0.01$ . This coupled model consists of three subsystems: an “*extratropical atmosphere*”, a “*tropical atmosphere*”, and a “*(tropical) ocean*”. Each component is simulated with a three-variable model of Lorenz (1963) but coupled by coefficients  $c$ ,  $c_z$ , and  $c_e$ . Furthermore, the “ocean” is slowed down by a factor of 10 through  $\tau$  to mimic the slower variation of the ocean. The extratropical atmosphere is only weakly coupled with the tropical atmosphere ( $c_e = 0.08$ ), and the tropical atmosphere is, in turn, strongly coupled with the ocean ( $c = c_z = 1$ ). There is no direct interaction between the extratropical atmosphere and the ocean.

Despite its extreme simplicity, the two-timescale coupled model shares important characteristics with the real coupled atmosphere-ocean system and is an excellent testbed for testing ideas for coupled DA problems. For example, the model shows a chaotic behavior with two distinct regimes: the coupled tropical atmosphere and the ocean cycle into a random number of “normal years” (between two and seven), interrupted by an anomalous “El Niño year”, before returning to normal years. Since this asymmetric oscillation neither occurs in the uncoupled tropical atmosphere nor ocean, it is regarded as an intrinsically coupled instability.

All the experiments with this model (results shown in subsection 2.4) are conducted with an LETKF cycle with an analysis window of 8 timesteps. Observations are only available at the end of each window, and only three of nine variables ( $y_e, y_t$ , and  $Y$ ) are observed unless otherwise noted. The ensemble size  $K$  for this nine-variable model with four non-negative Lyapunov exponents (Norwood et al., 2013) is chosen to be 4, 6, or 10, representing an insufficient-member, intermediate, and a sufficient-member case respectively.

## 2.4 Experimental results

First, we conducted offline experiments of LETKF to estimate the mean squared correlation of background error (5). For this purpose, a strongly coupled LETKF cycle with 10 members is used, and  $T = 6250$  analyses are computed. Figure 2 shows the mean squared correlation obtained with this experiment. In this model, there is a clear distinction between well-correlated ( $\langle \text{corr}^2 \rangle > 0.5$ ) pairs of variables and weakly correlated ( $\langle \text{corr}^2 \rangle < 0.03$ ) pairs of variables. Therefore, it is clear where we should stop assimilation: we should assimilate the observations in the “tropical atmosphere” and “ocean” together and analyze the “extratropical atmosphere” individually.

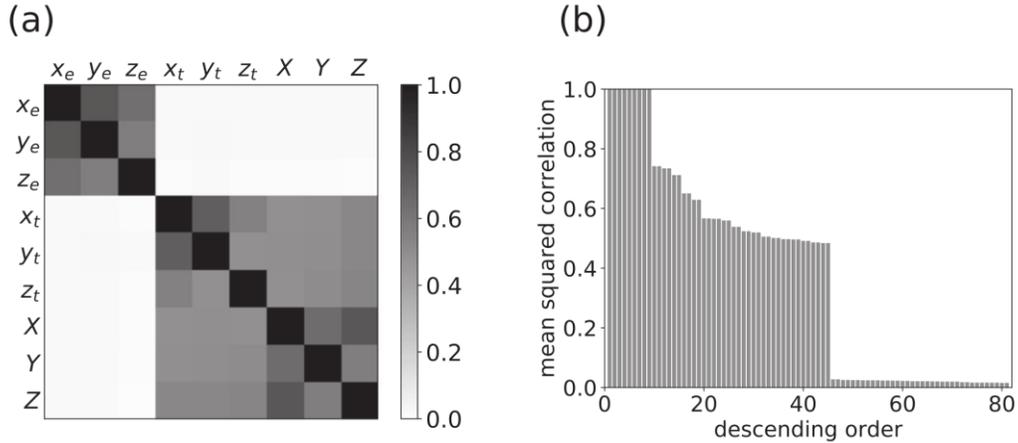


Figure 2: (a) Mean of squared error correlation for each pair of variables (5) obtained by an offline LETKF experiment with the nine-variable model of Peña and Kalnay (2004). (b) Mean of squared error correlation for all the 81 ordered pairs of variables, sorted in descending order. Note the drop in correlation after the first 45 highly correlated pairs of variables, which indicates that the *ENSO-coupling* localization pattern in Figure 4 is optimal. Figure from Yoshida and Kalnay (2018)

Figure 3 shows the result of other offline experiments with different observation networks and LETKF configurations. Although the distinction between the highly correlated and weakly correlated pair of variables becomes less apparent in the 4-member WCDA experiment (Figure 3b), we can see that the correlation structure of the background error is almost intrinsic in the model and not significantly affected by the observation network or the ensemble size. This robustness of the error correlation structure is supportive of the correlation-cutoff method since we do not need to re-evaluate the structure every time we change the observation or assimilation systems.

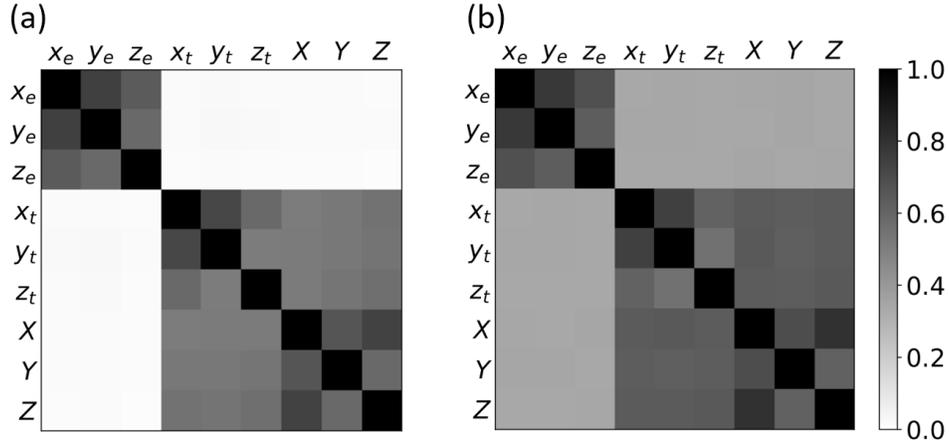


Figure 3: Same with Figure 2(a) but (a) estimated with an LETKF cycle assimilating observations of all nine variables instead of only  $y_e, y_t$ , and  $Y$  and (b) estimated with a 4-member WCDA cycle instead of a 10-member fully coupled SCDA cycle.

Then, we test five covariance localization patterns in Figure 4 and evaluate the analysis accuracy.

- The *Full* pattern is the standard SCDA, in which every observation is assimilated into every variable.
- In the *Adjacent* pattern, we ignore the cross-covariance between the extratropical atmosphere and the ocean, which are not directly interacting.
- The *ENSO-coupling* pattern is the one suggested by the theoretical analysis and the offline experiment (Figure 2a). Here, the tropical atmosphere and the ocean are mutually assimilated, but the extratropics is analyzed individually.
- In *Atmos-coupling*, we analyze the extratropical and tropical atmosphere at the same time but the ocean separately. This pattern separately analyzes the faster components and the slower component.

- The last is *Individual*, in which each component is analyzed individually. This pattern is equivalent to WCDA for this three-component model; the background is updated by the coupled model, but the analysis step is individually conducted in each component.

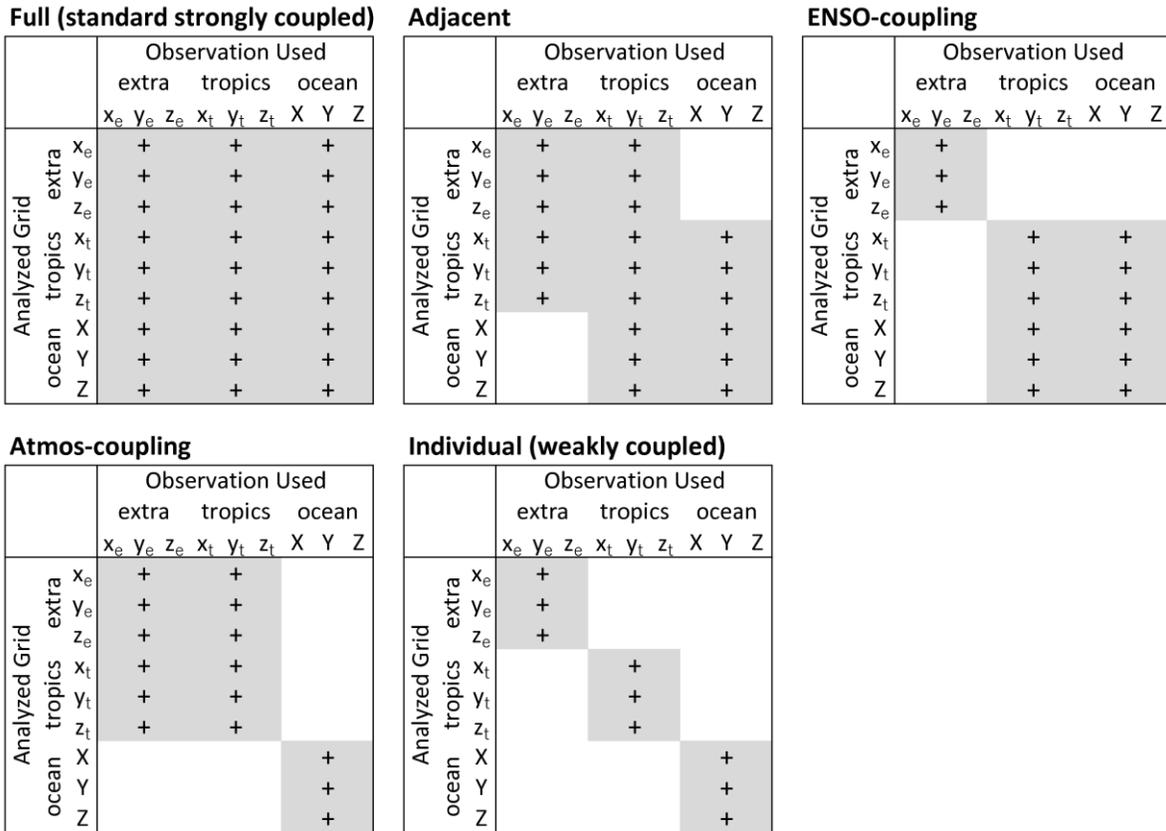


Figure 4: Covariance localization patterns tested. Background covariance is considered only at the shaded pairs of variables. Except for the offline experiment of Figure 3(a), only the observations (columns) indicated by “+” signs are assimilated into model variables (rows).

Figure from Yoshida and Kalnay (2018)

The resulting analysis errors are plotted in Figure 5.

We first note that *Full* (standard SCDA) performs worse than *Individual* (WCDA) when the ensemble size is small ( $K = 4, 6$ ). This negative result of considering the cross-covariance would be caused by the rank deficiency of the filter and the spurious correlation between the components. As the ensemble becomes larger, *Full* becomes gradually better, whereas the analysis accuracy of *Individual* is not so sensitive to the ensemble size. This result supports the importance of using larger ensembles for successful implementation of strongly coupled EnKFs as claimed by Han et al. (2013).

As Eq. (3) indicates, in the Kalman filter, the assimilation of any observation, on average, will not increase the analysis uncertainty if the assumptions on the error statistics are valid. Therefore, with an LETKF with sufficient ensemble members, assimilation of observations of uncorrelated variables will be neither beneficial nor harmful. The ensemble size needed for successful implementation of SCDA will be highly model dependent.

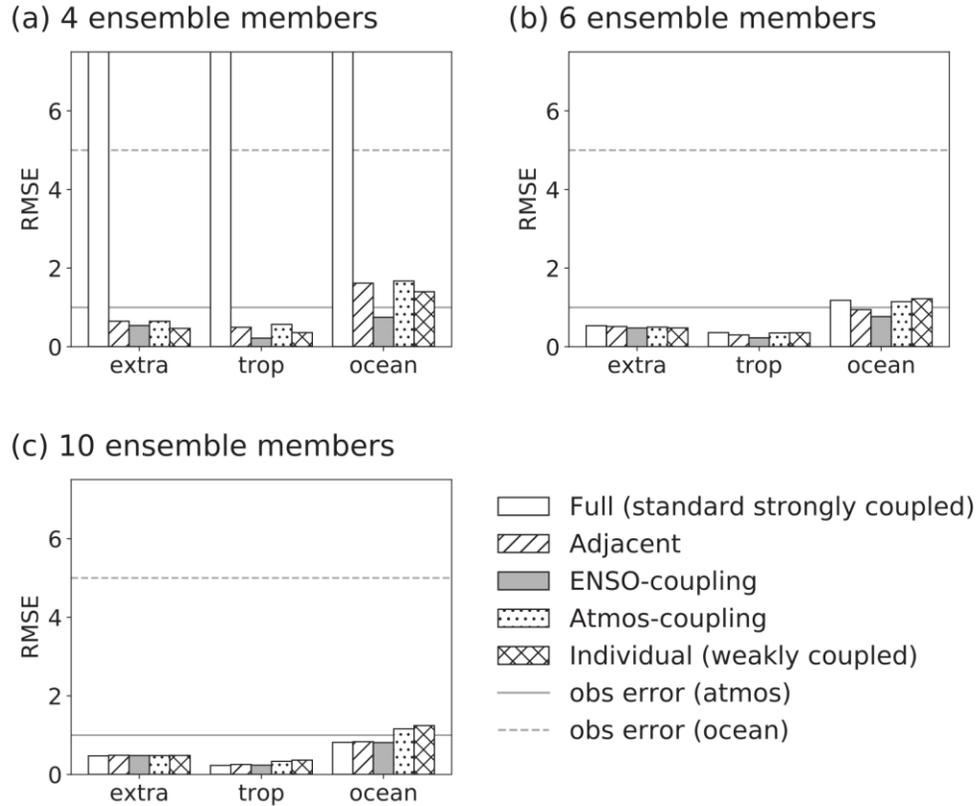


Figure 5: Temporal mean analysis root-mean-square error (RMSE) of each experiment. Different shading shows the different covariance localization pattern (see the legend and Figure 4). In each panel, the errors in the extratropical atmosphere, the tropical atmosphere, and the ocean are separately shown. Horizontal lines indicate the observation errors of each component for comparison. Each panel is the result of (a) 4-member experiments, (b) 6-member experiments, and (c) 10-member experiments. Note that the filter diverges and cannot finish all the cycle correctly in the 4-member Full experiment. Figure from Yoshida and Kalnay (2018)

The *ENSO-coupling* pattern suggested by the correlation-cutoff method performs best in essentially all experiments and components. In comparison to *Individual* and

*Atmos-coupling*, *ENSO-coupling* is superior regardless of the ensemble size. This result supports the importance of exploiting information from the strong background error correlation between the tropical atmosphere and the ocean in this model. In contrast to *Full* and *Adjacent*, *ENSO-coupling* performs well with the smaller ensembles ( $K = 4, 6$ ), while all these patterns, which consider the strong cross-covariance between the tropical atmosphere and the ocean, perform similarly well when the ensemble size is large enough ( $K = 10$ ). This result supports the decision of ignoring the cross-covariance between the extratropical atmosphere and the other components as suggested by Figure 2(a).

### 3. Experiments with a coupled general circulation model

#### 3.1 Experimental settings

We investigate the background error correlation of the Fast Ocean Atmosphere Model (FOAM; Jacob, 1997), to search for the optimal inter-fluid localization for SCDA. FOAM is a coupled atmosphere-ocean-sea-ice-land GCM with intermediate complexity. Table 1 shows its specifications. Although FOAM is a coupled GCM with physical parameterizations and can reproduce realistic climatology and internal variability of interest, it is implemented efficiently for parallel computing. When tested on the Deepthought2 supercomputer at the University of Maryland (Ivy Bridge, 2.8 GHz), it finished a 50-model-year integration within 3 hours with 16 processors. The efficiency of the model is essential for an early-stage study of coupled DA since EnKFs need to run tens of ensemble members in parallel. In addition, a coupled model has longer spin-up time than an atmospheric model, which means we need longer experiments to evaluate the method accurately.

The model, originally developed for the study of climate dynamics, has been slightly modified for the study of coupled DA. First, the model's shortest integration length (frequency of input and output) is shortened from 24 hours to 6 hours. This enables a 6-hourly DA cycle, which is typically adopted by most global atmospheric DA systems. Also, the 6-hour interval is as short as the model's coupling frequency between the atmosphere and the ocean, and it enables examining the coupled dynamics more closely. Second, the model is augmented with incremental analysis update (IAU; Bloom et al. 1996) so that the analysis increment is distributed into a finite time interval and gradually added

to the model state. The primary motivation for implementing IAU is to stabilize the atmospheric part of the coupled model, which often blows up due to numerical instability when the analysis increment is suddenly imposed.

For a DA algorithm, we adopt LETKF as we have done with the nine-variable coupled model. The LETKF program has been written to read/write the restart files of FOAM and analyzes horizontally divided domains in parallel using the Message Passing Interface (MPI). The LETKF program is designed specifically for exploring the coupled DA and deals with the atmospheric and oceanic variables in the same manner.

Table 1: Specification of the Fast Ocean Atmosphere Model (FOAM). See Chapter 2 of Jacob (1997) for details.

<b>Atmosphere model (PCCM3; spectral, hydrostatic)</b>	
Resolution	R15 spectral (40 latitudes $\times$ 48 longitudes) 18 levels (hybrid sigma-pressure coordinate)
Integration time step	30 minutes (radiation is only updated hourly)
Parameterized schemes	Convection Cloud Radiation ( $\delta$ -Eddington type) Vertical diffusion (boundary layer) Surface physics Gravity wave drag
<b>Ocean model (OM3; finite difference, hydrostatic, incompressible)</b>	
Resolution	128 latitudes $\times$ 128 longitudes (polar grid) 24 levels (z coordinate)
Integration time step	15 minutes (barotropic pressure balance) 2 hours (remaining pressure balance) 6 hours (advection and diffusion)
Parameterized schemes	Vertical mixing
<b>Coupler, Sea ice, Land, Hydrology, and River runoff</b>	
Resolution	128 latitudes $\times$ 128 longitudes (same as the ocean) The land model has four layers of soil
Atmosphere-ocean fluxes (freshwater, heat, and momentum)	Calculated in the atmosphere model using SST, accumulated in the coupler, and passed to the ocean model every six hours
Sea ice processes	Freezing/melting, freshwater exchange, albedo, surface roughness, and insulation
Integration time step	30 minutes

### 3.2 Background error correlation of a coupled GCM

After implementing the LETKF system for FOAM and roughly tuning the DA parameters such as inflation and (distance-based) localization, we thoroughly examined the ensemble correlation of the coupled system; this mirrors the offline experiments with the nine-variable model (i.e., Figure 2 and Figure 3). All results in this subsection are from weakly-coupled, 64-member observation system simulation experiments (OSSEs).

In this subsection, the temporal mean ensemble correlation instead of the temporal mean *squared* ensemble correlation (3) is shown. One reason is that the former is easier to interpret by associating with dynamical processes as it retains positive and negative signs. Another reason is that the square of the mean correlation ( $\langle \text{corr} \rangle^2$ ) and the mean squared correlation ( $\langle \text{corr}^2 \rangle$ ) do not qualitatively differ in our examples. Note that only the former includes the temporal fluctuation (denoted by a prime):  $\langle \text{corr}^2 \rangle = \langle \text{corr} \rangle^2 + \langle \text{corr}'^2 \rangle$ . This means that the flow-dependent portion of the background error correlation is not dominating in our examples.

First, vertical ensemble correlation structure between each pair of atmospheric ( $U$ ,  $V$ ,  $T$ ,  $Q$ ,  $P_s$ ) and oceanic ( $U$ ,  $V$ ,  $T$ ,  $S$ ) state variables is examined at several points over the ocean. Two examples that have strong cross-correlation are shown in Figure 6.

Figure 6(a) is mean ensemble correlation between the meridional wind and the zonal current at the southern Indian Ocean. The atmosphere, especially the troposphere and the lower stratosphere is dominated by barotropic error structure, with positive error correlation between those layers. The internal ocean below 50 meters similarly has barotropic error structure with positive error correlation between layers. The background error of zonal

current within a few upper layers ( $< 50$  m depth) of the ocean is less correlated to that of the internal ocean and negatively correlated with the meridional wind error. It is known that in the southern hemisphere, the water in the top Ekman layer is transported to the left of the wind stress so that the Coriolis force of the transported water balances the wind stress (e.g., Figure 2.14 of Vallis, 2006). Under a first-order assumption that the wind stress is proportional to the wind velocity and that the oceanic eddy viscosity coefficient to be independent of the wind stress, the wind and Ekman transport have a linear relationship; hence the *errors* of wind and current also follow the Ekman transport relationship. Such background error correlation between the surface wind and the perpendicular surface current is seen everywhere in the midlatitudes and strongest among the cross-correlations between the atmosphere and the ocean. The spatial and temporal average surface correlation between errors of the meridional wind and the zonal current reaches 0.76 (-0.79) north of  $20^{\circ}\text{N}$  (south of  $20^{\circ}\text{S}$ ).

Figure 6(b) is mean ensemble correlation between the temperatures of the atmosphere and the ocean, at the Eastern Equatorial Pacific, where the anomalous SST is expected to drive the atmosphere. Within the atmosphere, the errors of temperature in each layer are almost independent of each other except in the planetary boundary layer at the bottom. In the planetary boundary layer, there is positive error correlation within the adjacent two or three layers. Within the ocean, above 1000 m depth, the temperature error of each layer is loosely and positively correlated with each other. Below 1000 m, the temperature error shows almost perfect internal correlation and is slightly negatively correlated to that of the upper ( $< 1000$  m depth) layers. The cross-correlation between the upper ocean and the lower atmosphere is positive, as expected. Within  $20^{\circ}\text{S}$ - $20^{\circ}\text{N}$ , the spatial and temporal

average cross-correlation between the surface temperatures is 0.41, which means almost 17% ( $0.41^2$ ) of the error variance of surface air temperature can be explained by the error variance of sea surface temperature, and vice versa.

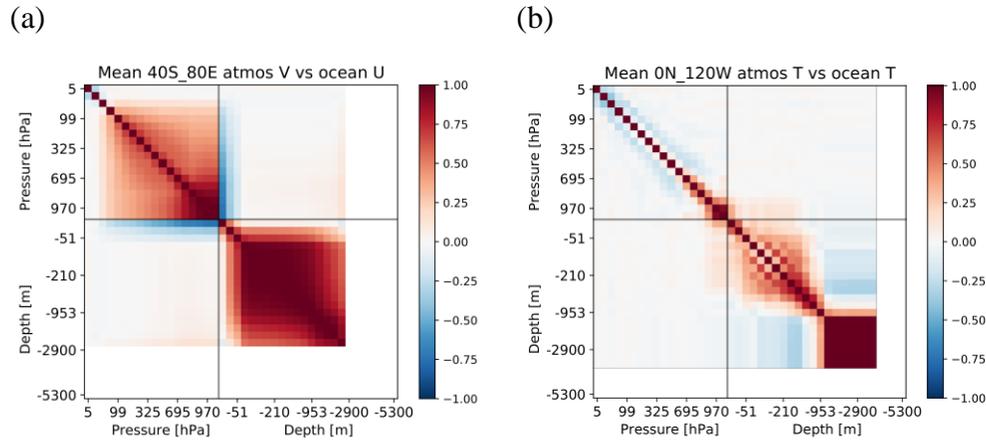


Figure 6: Ensemble correlation within vertical columns averaged for 260 instances (00UTC of 260 days starting on model date January 1<sup>st</sup>). The panels show (a) ensemble correlation between meridional wind and zonal current at point 40°S 80°E (southern Indian Ocean) and (b) ensemble correlation between atmospheric and oceanic temperature at 0° 120°W (Eastern Equatorial Pacific). Deep ocean layers below the model topography are whited-out. Upper-left and lower-right quadrants of each panel show atmospheric and oceanic internal error correlation, respectively, and the other two quadrants show cross-correlation between the atmosphere and the ocean.

It is also informative to check the horizontal structure of error correlation to understand the driving mechanisms of error growth in the coupled system. Figure 7 shows temporal mean background ensemble correlation to an observation background.

Figure 7(a) shows a map of background error correlation between the surface zonal current of each grid and a surface zonal current observation at  $40^{\circ}\text{S}$   $80^{\circ}\text{E}$  (southern Indian Ocean). Naturally, the peak of positive correlation exists around the observation location. In addition to the main lobe of correlation, the error has two lobes of negative correlation to the east and west of the observation. Interestingly, the error correlation in the ocean part of the coupled model has correlation extending to a continental scale; in other words, the oceanic error correlation extends to the synoptic scale of the atmosphere. This large-scale error correlation in the ocean cannot be explained by the ocean's internal dynamics because the ocean is thought to have small decorrelation length; typical middle-latitude ocean has Rossby deformation radius of 100 km (e.g., section 5.2 of Vallis, 2006).

If we look at the background error correlation of surface wind, we can understand the cause of the large-scale error correlation in the ocean. Figure 7(b) is a map of background error correlation between the surface meridional wind of each atmospheric grid and an observation of surface meridional wind at  $40^{\circ}\text{S}$   $80^{\circ}\text{E}$ . Although Figure 7(a) and (b) respectively show the error correlation structure within the ocean and the atmosphere, they look almost identical. The spatial structure of Figure 7(b) resembles the spatial correlation function under the geostrophic increment assumption (Figure 5.4.2 of Kalnay, 2003). Therefore, we can explain that the continental-scale correlation originates from the atmospheric quasi-geostrophic dynamics. At each point, the surface ocean current is driven by the atmospheric wind, and so is its error.

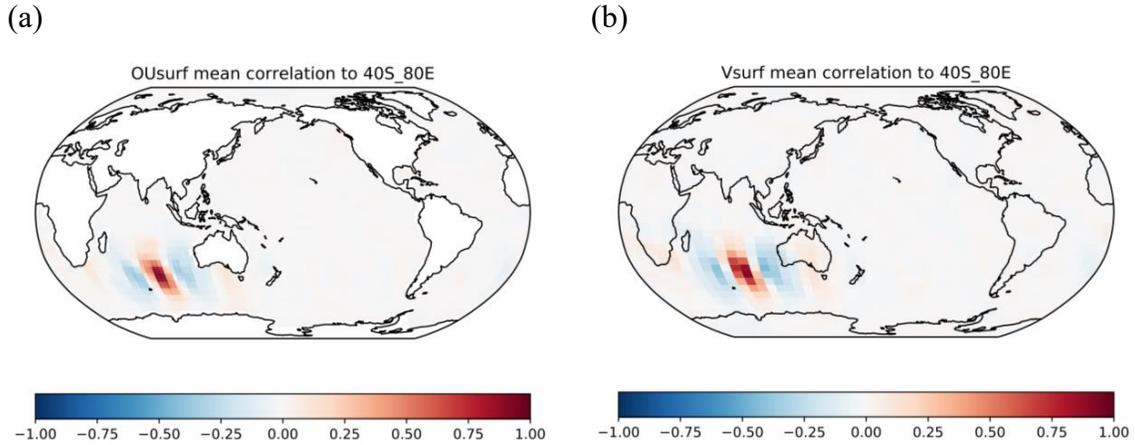


Figure 7: Mean ensemble background correlation of (a) ocean surface zonal current to an ocean surface zonal current observation at point 40°S 80°E (southern Indian Ocean) and (b) atmospheric surface meridional wind to an atmospheric surface meridional wind observation at the same point, averaged for 260 instances (00UTC of 260 days starting on model date January 1<sup>st</sup>).

These examples show the existence of physically reasonable cross-correlation between the atmosphere and the ocean and support the possibility of improving the coupled DA by utilizing the cross-covariances. However, it should be noted that the spatial and temporal scale of those cross-correlation varies by latitudes and variables. For example, the ensemble correlation between SST and surface air temperature becomes greater when atmospheric ensemble perturbation is averaged for several days (as done by Lu et al., 2015b), whereas the error correlation between surface current and surface wind weakens when temporally smoothed out (not shown). Therefore, it is essential to understand the characteristics of each coupling process for the SCDA applications.

#### 4. Summary and Future Work

In this study, we first compared weakly and strongly coupled DA in light of background error covariance/correlation used in the analysis step of data assimilation. Here we showed that the uncertainty reduction by the assimilation of an observation is proportional to the square of background error correlation between the observed and the analyzed variables. From this observation, we proposed the correlation-cutoff method, where the analysis step of a coupled system should be coupled only between the well-correlated pair of variables in term of background error.

The correlation-cutoff method was tested with a nine-variable coupled model which mimics the fast and slow variations of the atmosphere-ocean coupled system. With the model, partially coupled analysis, guided by the offline error statistics, achieved the best accuracy out of five coupling (localization) patterns tested. The analysis accuracy of the correlation-cutoff method was more robust than the standard strongly coupled DA for smaller ensemble sizes.

Finally, to apply the correlation-cutoff method to strongly coupled DA of the atmosphere-ocean system, the background error correlation of a coupled GCM (FOAM) was examined. With the coupled model, the strongest background error correlation was found in the Ekman layer dynamics in the midlatitudes. The ocean surface current, which is thought to be the ‘slave’ of the wind in the midlatitudes, was found to have continental-scale horizontal error correlation originated from the atmospheric quasi-geostrophic dynamics.

The correlation-cutoff method will be extended to the localization of the global strongly coupled analysis of the atmosphere and ocean. The next important step will be to summarize the mean squared background error correlation between the observed and the analyzed variables into a simple function of their distance and variable types. We will test the methodology with the same analysis system for the coupled GCM, which should be an important milestone in the design of an improved strongly coupled DA.

## References

- Argo (2000). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE.
- Bloom, S. C., L. L. Takacs, A. M. Da Silva, and D. Ledvina (1996). Data Assimilation Using Incremental Analysis Updates. *Mon. Weather Rev.*, **124**, 1256–1271.
- Chen, D., Zebiak, S. E., Busalacchi, A. J., and Cane, M. A. (1995). An Improved Procedure for El Niño Forecasting: Implications for Predictability. *Science*, **269**, 1699–1702.
- Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162.
- Frolov, S., Bishop, C. H., Holt, T., Cummings, J., and Kuhl, D. (2016). Facilitating Strongly Coupled Ocean-Atmosphere Data Assimilation with an Interface Solver. *Mon. Wea. Rev.*, **144**, 3–20.
- Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., and Hunt, B. R. (2011). Balance and Ensemble Kalman Filter Localization Techniques. *Mon. Wea. Rev.*, **139**, 511–522.
- Hamill, T. M., J. S. Whitaker, and C. Snyder (2001). Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. *Mon. Weather Rev.*, **129**, 2776–2790.

- Han, G., Wu, X., Zhang, S., Liu, Z., and Li, W. (2013). Error Covariance Estimation for Coupled Data Assimilation Using a Lorenz Atmosphere and a Simple Pycnocline Ocean Model. *J. Climate*, **26**, 10218–10231.
- Houtekamer, P. L. and Mitchell., H. L. (2001). A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. *Mon. Wea. Rev.*, **129**, 123–37.
- Hunt, B. R., Kostelich, E. J., and Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.
- Jacob, R. L. (1997). Low Frequency Variability in a Simulated Atmosphere Ocean System. PhD Thesis, 159pp.
- Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.*, **82**, 35–45.
- Kalnay, E. (2003). Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, 341pp.
- Kang, J.-S., Kalnay, E., Liu, J., Fung, I., Miyoshi, T., and Ide, K. (2011). “Variable localization” in an ensemble Kalman filter: Application to the carbon cycle data assimilation. *J. Geophys. Res.*, **116**, D09110.
- Karspeck, A. R., and Coauthors (2018). A global coupled ensemble data assimilation system using the Community Earth System Model and the Data Assimilation Research Testbed. *Q. J. R. Meteorol. Soc.*, Preprint, doi:10.1002/qj.3308.

- Kotsuki, S., S. J. Greybush, and T. Miyoshi (2017). Can We Optimize the Assimilation Order in the Serial Ensemble Kalman Filter? A Study with the Lorenz-96 Model. *Mon. Weather Rev.*, **145**, 4977–4995.
- Laloyaux, P., M. Balmaseda, D. Dee, K. Mogensen, and P. Janssen (2016). A coupled data assimilation system for climate reanalysis. *Q. J. R. Meteorol. Soc.*, **142**, 65–78.
- Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, **20**, 130–141.
- Lu, F., Liu, Z., Zhang, S., and Liu, Y. (2015a). Strongly Coupled Data Assimilation Using Leading Averaged Coupled Covariance (LACC). Part I: Simple Model Study\*. *Mon. Wea. Rev.*, **143**, 3823–3837.
- Lu, F., Liu, Z., Zhang, S., Liu, Y., and Jacob, R. (2015b). Strongly Coupled Data Assimilation Using Leading Averaged Coupled Covariance (LACC). Part II: CGCM Experiments\*. *Mon. Wea. Rev.*, **143**, 4645–4659.
- McPhaden, M. J., Busalacchi, A. J., Cheney, R., Donguy, J.-R., Gage, K. S., Halpern, D., Ji, M., Julian, P., Meyers, G., Mitchum, G. T., Niiler, P. P., Picaut, J., Reynolds, R. W., Smith, N., and Takeuchi, K. (1998). The Tropical Ocean-Global Atmosphere observing system: A decade of progress. *J. Geophys. Res.*, **103**, 14169–14240.
- Mulholland, D. P., Laloyaux, P., Haines, K., & Balmaseda, M. A. (2015). Origin and Impact of Initialization Shocks in Coupled Atmosphere-Ocean Forecasts\*. *Mon. Wea. Rev.*, **143**, 4631–4644.
- Nerger, L. (2015). On Serial Observation Processing in Localized Ensemble Kalman Filters. *Mon. Weather Rev.*, **143**, 1554–1567.

- Norwood, A., Kalnay, E., Ide, K., Yang, S.-C., and Wolfe, C. (2013). Lyapunov, singular and bred vectors in a multi-scale system: an empirical exploration of vectors related to instabilities. *J. Phys. A*, **46**, 254021.
- Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D., and Yorke, J. A. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, **56**, 415–428.
- Peña, M. and Kalnay, E. (2004). Separating fast and slow modes in coupled chaotic systems. *Nonlinear Process Geophys.*, **11**, 319–327
- Penny, S. G. and Hamill, T. M. (2017). Coupled Data Assimilation for Integrated Earth System Analysis and Prediction. *Bull. Amer. Meteor. Soc.*, **98**, ES169-ES172.
- Penny, S. G., and Coauthors. (2017). Coupled Data Assimilation for Integrated Earth System Analysis and Prediction: Goals, Challenges and Recommendations. World Meteorological Organization, 50pp.
- Ruiz-Barradas, A., E. Kalnay, M. Peña, A. E. BozorgMagham, and S. Motesharrei (2017). Finding the driver of local ocean–atmosphere coupling in reanalyses and CMIP5 climate models. *Clim. Dyn.*, **48**, 2153–2172.
- Saha, S., and Coauthors. (2010). The NCEP Climate Forecast System Reanalysis. *Bull. Am. Meteorol. Soc.*, **91**, 1015–1057.
- Sluka, T. C., Penny, S. G., Kalnay, E., and Miyoshi, T. (2016). Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation. *Geophys. Res. Lett.*, **43**, 752–759.

- Smith, P. J., Lawless, A. S., and Nichols, N. K. (2017). Estimating Forecast Error Covariances for Strongly Coupled Atmosphere-Ocean 4D-Var Data Assimilation. *Mon. Wea. Rev.*, **145**, 4011–4035.
- Vallis, G. K. (2006). *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*. 1st ed. Cambridge University Press, 745 pp.
- Yoshida, T., and E. Kalnay (2018). Correlation-Cutoff Method for Covariance Localization in Strongly Coupled Data Assimilation. *Mon. Weather Rev.*, **146**, 2881–2889.
- Zhang, S., Harrison, M. J., Rosati, A., and Wittenberg, A. (2007). System Design and Evaluation of Coupled Ensemble Data Assimilation for Global Oceanic Climate Studies. *Mon. Wea. Rev.*, **135**, 3541–3564.
- Zhang, S., Rosati, A., and Delworth, T. (2010). The Adequacy of Observing Systems in Monitoring the Atlantic Meridional Overturning Circulation and North Atlantic Climate. *J. Climate*, **23**, 5311–5324.