RMetS
Royal Meteorological Society

Growing-error correction of ensemble Kalman filter using empirical singular vectors

Yoo-Geun Ham[a,b] and In-Sik Kang[c]*
[a]*GSFC Global Modeling and Assimilation Office, Greenbelt, Maryland, USA*
[b]*Goddard Earth Sciences and Technology Center, University of Maryland, Baltimore, Maryland, USA*
[c]*School of Earth and Environment Sciences, Seoul National University, South Korea*
*Correspondence to: In-Sik Kang, Seoul National University 501-420, SNU Sillim-Dong Seoul 151-742 South Korea.
E-mail: kang@climate.snu.ac.kr

In this study, a new Ensemble Kalman Filter (EnKF) algorithm called EnKF with growing-error correction (EnKF-GEC) is developed for minimizing the growing component of the forecast error; for this purpose, prospective observations are assimilated using empirical singular vectors (ESVs). Unlike the Ensemble Kalman Smoother (EnKS) or four-dimensional EnKF (4DEnKF), the EnKF-GEC is designed to reduce the analysis error at the last analysis time (errors of initial condition for prediction). By performing assimilation experiments using the CZ-SPEEDY coupled model within a perfect model framework, it is shown that the analysis errors obtained using the EnKF-GEC are significantly reduced as compared to those obtained using the conventional EnKF until the last analysis time as well as during the middle of analysis time. This indicates that the new algorithm is beneficial for prediction. Seasonal prediction results show that the prediction skill when initial conditions are generated by the EnKF-GEC is superior to when initial conditions are generated by the conventional EnKF or EnKS, particularly during the early forecast lead month. For example, correlation skill improvement with 16 ensemble members is about 0.1 for a 3-month lead forecast. In addition, it is shown that the new EnKF algorithm is more effective for unpredictable regions, where the value of the unstable singular vector is robust. Copyright © 2010 Royal Meteorological Society

Key Words: Ensemble Kalman Filter; Empirical Singular Vector; unstable modes

Received 3 March 2010; Revised 10 September 2010; Accepted 14 September 2010; Published online in Wiley Online Library 29 October 2010

Citation: Ham Y-G, Kang I-S. 2010. Growing-error correction of ensemble Kalman filter using empirical singular vectors. *Q. J. R. Meteorol. Soc.* **136**: 2051–2060. DOI:10.1002/qj.711

1. Introduction

The Ensemble Kalman Filter (EnKF) was introduced for applications in oceanic and atmospheric sciences, and the data assimilation community developed the EnKF algorithm using two approaches. One approach is to obtain an accurate background error covariance matrix estimated using ensemble spread within the EnKF framework, e.g. Gaussian probability distribution function (pdf) assumptions with linear dynamics (Burgers *et al.*, 1998; Anderson and Anderson, 1999; Evensen, 2003; Hunt *et al.*, 2004). In this framework, methods to obtain accurate background model error covariance by adding random perturbation to the analysis ensemble perturbation, or applying covariance inflation, are developed. The other approach involves assimilation of observations at times that are different from the assimilation time (Cohn *et al.*, 1994; Evensen and van Leeuwen, 2000; Huang *et al.*, 2002; Hunt *et al.*, 2004). Many researchers working with this framework expect that with the use of future data, in addition to current and past data, the effective amount of data available for each analysis will be doubled, and analysis errors will be reduced.

One of the simple approaches to assimilate the observations at different times is the First Guess at Appropriate Time (FGAT) approach, which was mentioned in Lorenc (2003). FGAT takes the innovation (difference between the observations and the model forecasts) at an asynchronous time, and treats this innovation as if it is calculated at the analysis time (Huang *et al.*, 2002; Lorenc and Rawlins, 2005). However, as expected, this method ignores the time difference of the background error covariance between the prospective innovations and model forecasts at the analysis time; therefore, a more accurate algorithm is required for treating future observations and for obtaining optimal estimates using asynchronous observations.

To develop the above-mentioned algorithm, Cohn *et al.* (1994) introduced the Fixed-Lag Kalman Smoother (FLKS) for retrospective data assimilation using future observations. FLKS is designed to use the temporal cross-covariance for calculating the weighting coefficients of the innovation (observations minus first guess) at the prospective time to perform an updated retrospective analysis. Cohn *et al.* (1994) showed that the FLKS is capable of successfully propagating information upstream as well as downstream, therefore significantly improving the analysis quality, using a two-dimensional linear shallow-water model. After their initial attempts, Evensen and van Leeuwen (2000) extended the Kalman smoother to ensemble-based initialization referred to as the Ensemble Kalman Smoother (EnKS), and they proved that in a simple Lorenz system, the analysis quality of the EnKS is superior to that of the EnKF.

Hunt *et al.* (2004) developed a four-dimensional ensemble Kalman filtering algorithm to unify the Kalman Filter and the four-dimensional variational method (4D-Var). Their key idea was that the analysis value is a linear combination of the ensemble forecasts which best fits the observations within the assimilation window, and the analysis equation was modified to find the weighting coefficients for each forecast ensemble. Therefore, the cost function is formulated with respect to the weighting coefficient of each ensemble, and no linear adjoint operator is required for minimizing the cost functions. The application of 4DEnKF to the Lorenz-96 model and SPEEDY atmospheric general circulation model (AGCM) (simplified parametrizations, primitive-equation dynamics: Molteni, 2003) results in improved analysis quality and is better than that achieved with 4D-Var and FGAT (Fertig *et al.*, 2007; Harlim and Hunt, 2007).

There are also several attempts to utilize the retrospective observations using a particle filter. For example, the auxiliary particle filter introduced by Pitt and Shephard (1999) determines the first-stage weighting of each ensemble member using the probability density function, then re-runs the forecast model. They expected that this re-sampled ensemble would provide forecasts closer to the observations.

However, most studies pertaining to the assimilation of future observations have been performed within the analysis framework. This implies that the importance of analysis quality at the last analysis time (initial time of prediction) has been underestimated until now, although improved analysis quality at the last analysis time is required for improving the prediction capability of the algorithm. Therefore, the development of the data assimilation algorithm has not been directly related to the improvement of the prediction quality. For example, the estimate at the last analysis time is identical for the EnKS and EnKF (Evensen and van Leeuwen, 2000). In addition, the result of the 4D-Var analysis performed at the end of the assimilation window is equal to that of the Kalman filter analysis performed under a linear assumption. By laying emphasis on this point, we developed the EnKF algorithm to reduce the analysis error at the last analysis time by removing growing errors in the analysis values.

This paper is organized as follows. In section 2, a new algorithm (called EnKF-GEC) for eliminating growing errors and its comparison to the EnKS are described. In section 3, descriptions of the forecast model and explanations of the experimental design are mentioned. In section 4, the data assimilation results and seasonal prediction results obtained using the EnKF-GEC algorithm are described and compared with those obtained using the conventional EnKS. A summary of the study and discussions are included in section 5.

2. EnKF with Growing-Error Correction (EnKF-GEC), and comparison with the Ensemble Kalman Smoother (EnKS)

2.1. *Description of EnKF-GEC*

Figure 1 shows the schematic diagram of the EnKF algorithm with growing-error correction (EnKF-GEC). The analysis time (target time) is assumed to be T. After using the conventional EnKF for obtaining the analysis states for each ensemble at time T, free integration from time T to $T+1$ is performed using these analysis states. The integrated states are denoted as model forecasts at time $T+1$. Using the model forecasts and observations at time $T+1$, the analysis states at time $T+1$ are obtained by applying the conventional EnKF again. It should be noted that according to data assimilation theory, the analysis states are more accurate than both observations and model forecasts; the analysis states are defined as approximated true states (Gelaro and Zhu, 2009). Therefore, deviations of model forecasts from the analysis states are defined as approximated forecast errors at time $T+1$ ($E(T+1)$). Then, the defined forecast errors are projected onto the target assimilation time using an empirical singular vector (ESV) whose initial (final) time is T ($T+1$) (Kug *et al.*, 2009). The difference of this study from Kug *et al.* (2009) is that ensemble perturbations are utilized for each sample.

This method is based on the concept of Singular Vector Method (Farrell, 1989; Palmer *et al.*, 1994), but it uses an empirical operator instead of a dynamical operator. First, let us assume that nonlinear integration can be approximately expressed by a simple linear operator ($\mathbf{L}$) of the evolution of the state vector from time T to time $T+1$ as follows:

$$\mathrm{F_T} = \mathrm{I_{T+1}} = \mathrm{LI_T} + \in$$

where $\mathbf{I}$ and $\mathbf{F}$ are ensemble perturbations at time T and $T+1$. Note that the time difference between times T and $T+1$ is 1 month in this study. The variable $\in$ denotes error from the linear approximation. In the conventional singular vector method, the linear operator, $\mathbf{L}$, is calculated by linearizing the governing equation of the prediction model. However, it is also possible to estimate the operator empirically from ensemble samples. Then, the linear operator can be calculated as follows:

$$\mathrm{L} = \mathrm{FI}^{\mathrm{T}}(\mathrm{II}^{\mathrm{T}})^{-1}.$$

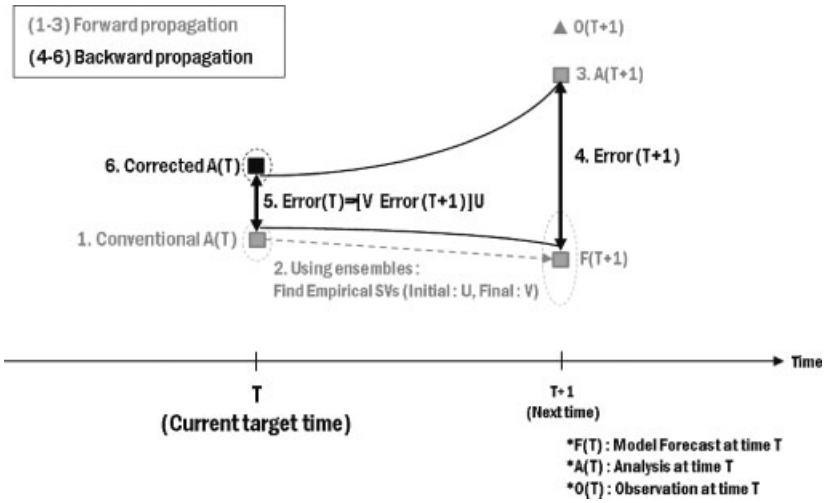

Figure 1. Schematic diagram of the EnKF algorithm with growing-error correction (EnKF-GEC).

This formula is basically the same as the linear inverse modelling approach (e.g. Blumenthal, 1991; Penland, 1996; Moore and Kleeman, 2001). Note that the dimension of **L** is n by n, when the number of ensemble members is n. By solving singular values of the linear operator, **L**, the singular vectors can be calculated:

$$u_i F = s_i v_i I$$

where s_i, u_i and v_i are the ith singular value and its corresponding singular vectors, respectively. If the singular value, s_i, is greater than unity, it implies that the singular mode grows in the linear operator.

The advantage of the ESV method is that there is a one-to-one relationship between the initial and final singular vectors, as in the case of singular vector (SV) methods. In addition, singular vectors are orthogonal to each other. This implies that each initial singular vector has only one corresponding final singular vector. Therefore, the approximated forecast errors at time $T+1$ ($E(T+1)$) can be propagated to time T by using the ESV, as follows:

$$E(T) = \sum_{n=1}^{m} [E(T+1) \cdot Y_n] X_n.$$

X_n and Y_n are the n^{th} normalized initial and final ESVs, respectively. m is the number of growing SVs. This procedure shows that $E(T)$ is obtained by calculating the projection of the $E(T+1)$ pattern onto each of the final ESV modes. Then, the pattern projection coefficients are multiplied by the corresponding initial ESV at time T. Note that only growing singular modes, whose singular value is greater than 1, are used to carry out a corrected analysis; this is because a limited number of ensemble members result in sampling errors that in turn lead to pseudo-generation of minor singular modes, which are not physical modes. Also this prevents overweighting of observations to analysis values. This will be discussed later. Finally, the approximated forecast error at time T is subtracted from the conventional analysis values to negate the forecast errors:

$$A_{GEC}(T) = A(T) - E(T).$$

$\mathbf{A}_{GEC}$ and **A** are the analysis values obtained using the EnKF-GEC and the conventional EnKF, respectively. The corrected analysis of each ensemble member is freely integrated from time T to $T+1$, and then the same procedure is repeated up to the last analysis time.

The summary of the EnKF-GEC procedures is as follows. Note that the number of the procedure is the same as that shown in Figure 1.

(1) Perform conventional EnKF by using observations, and model forecasts at time T.
(2) Free integration from time T to $T+1$ is performed with these analysis values. Then, calculate empirical singular vectors using ensemble perturbations at time T and $T+1$.
(3) Perform conventional EnKF by using observations, and model forecasts at time $T+1$.
(4) Define approximated forecast errors at time $T+1$ as a difference of model forecasts from analysis values.
(5) Approximated forecast errors at time $T+1$ are propagated backward in time using ESVs, and are defined as approximated forecast errors at time T.
(6) Approximated forecast error at time T is subtracted from conventional analysis values to obtain final corrected analysis of EnKF-GEC.

Note that the analysis procedure in the case of the EnKF-GEC at the last analysis time is the same as that for the conventional EnKF because future observations are not available. However, an analysis value obtained with the EnKF-GEC at the last analysis time is still expected to be more accurate than that obtained with the conventional EnKF, because the new algorithm reduces the growing error in the analysis value calculated in the previous time interval. Therefore, the model forecast (background value) at the last analysis time is expected to be more accurate than that of conventional EnKF.

One can argue that in the new algorithm, observations are used twice for calculating each estimate, hence this appears to violate the rule that each observation should be used only once to obtain the analysis values. For example, the observations at time $T+1$ are used twice to estimate the approximated forecast error at $T+1$ ($(E(T+1))$), and conventional EnKF analysis states when the analysis time (target time) is $T+1$. However, observations can be used as much as necessary without any possibility of overweighting of observations (Kalnay and Yang, 2010; Yang and Kalnay,

2010). For example, the auxiliary particle filter (Pitt and Shephard, 1999) uses observations twice; to determine the first-stage weights, and to re-weight the ensemble members. This might look like overfitting to the observations; however, it is not because the multiplication of first-stage weightings and final weighting is theoretically equivalent to that of weighting of conventional particle filters. To prevent an overweighting problem of the observations, the new algorithm is designed to use each observation twice only when the analysis value obtained with the conventional EnKF is considered to be sub-optimal. It is done by the procedure that additional correction is performed when there are robust singular modes, which leads analysis errors (analysis minus true states) to grow faster. Note that the observations are used once when all the singular modes are stable (i.e. singular value is smaller than 1).

In addition, the observations used for estimating the approximated forecast error only affect the magnitude of the corrected error ($E(T)$), and the spatial pattern of the correction term. The spatial pattern of the corrected error at the analysis time is a linear combination of growing ESVs. Therefore, the authors believe that it is unlikely that the analysis will be overfitted to the observations.

2.2. Comparison of EnKF-GEC with the Ensemble Kalman Smoother (EnKS)

2.2.1. Comparison of analysis procedures to assimilate prospective observations

In this subsection, the difference between EnKF-GEC and EnKS in assimilating prospective observations is discussed. For a simple comparison, some notation is adopted from Cohn *et al.* (1994). $X^a_{(0|n)}$ means the analysis values resulting from the assimilation of an observation y_n at a time n into a background state at time 0. In addition, $M_{n\leftarrow 0}$ is a forecast model that transforms the initial model state X_0 into a future state X_n. Assume that the observation operator H is independent of time, and the assimilation window is from initial time T to future time $T+1$, and only one future observation at time $T+1$ is used to assimilate an analysis value at initial time (X_T).

The EnKS solution at time T after assimilating a retrospective observation at time $T+1$ is:

$$X\_EnKS^a_{(T|T+1)} = X^a_{(T|T)} + [M_{T+1\leftarrow T}P^a_{(T|T)}]^T H^T [HM_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}H^T + R]^{-1}[y_{T+1} - H(M_{T+1\leftarrow T}X^a_{(T|T)})]$$

Note that the covariance ($P^a_{(T|T)}$) in the ensemble framework is $[X'^a_{(T|T)}][X'^a_{(T|T)}]^T$. Note that the term $M_{T+1\leftarrow T}P^a_{(T|T)}$ can be written as the ensemble covariance between time T and $T+1$ (e.g. $[X'^b_{(T+1|T)}][X'^a_{(T|T)}]^T$).

Meanwhile, the procedure of EnKF-GEC propagates approximated forecast errors (defined as difference between forecast and analysis values at time $T+1$) backward in time. The analysis values at time $T+1$ ($X^a_{(T+1|T+1)}$) can be denoted as:

$$X^a_{(T+1|T+1)} = M_{T+1\leftarrow T}X^a_{(T|T)} + [M_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}]^T H^T [HM_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}H^T + R]^{-1}[y_{T+1} - H(M_{T+1\leftarrow T}X^a_{(T|T)})]$$

Note that $M_{T+1\leftarrow T}X^a_{(T|T)}$ and $M_{T+1\leftarrow T}P^a_{(T|T)}M_{T+1\leftarrow T}$ are forecast values and the corresponding error covariance matrix at time $T+1$, respectively. Then the approximated forecast error at time $T+1$ ($E(T+1)$) is defined as the difference between forecast ($M_{T+1\leftarrow T}X^a_{(T|T)}$) and analysis values:

$$E(T+1) = -[M_{T+1\leftarrow T}P^a_{(T|T)}M_{T+1\leftarrow T}]^T H^T [HM_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}H^T + R]^{-1}[y_{T+1} - H(M_{T+1\leftarrow T}X^a_{(T|T)})]$$

The backward propagation of the error information in time is using the growing modes of the empirical singular vectors (ESVs). Let the backward propagator using ESVs from $T+1$ to T be denoted as $M^{ESV}_{T\leftarrow T+1}$, then the analysis values for EnKF-GEC is

$$X\_GEC^a_{(T|T+1)} = X^a_{(T|T)} + [M^{ESV}_{T\leftarrow T+1}][M_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}]^T H^T [HM_{T+1\leftarrow T}P^a_{(T|T)}M^T_{T+1\leftarrow T}H^T + R]^{-1}[y_{T+1} - H(M_{T+1\leftarrow T}X^a_{(T|T)})]$$

It implies that, if the linear propagator is the same as the empirical operators with all singular modes being used (e.g. $[M^{ESV}_{T\leftarrow T+1}][M_{T+1\leftarrow T}] = I$), the analysis solution of EnKF-GEC and EnKS is equivalent. In addition, if there are no growing SVs, the solution of EnKF-GEC is equivalent to that of EnKF (e.g. $[M^{ESV}_{T\leftarrow T+1}] = 0$).

As is well known, EnKS is the optimal smoother solution for linear problems with Gaussian statistics. Similarly, the algorithms of EnKF-GEC and EnKF assume linear dynamics and Gaussian pdf. This assumption would not give the correct solution if the prior joint pdf has non-Gaussian contributions. To overcome this deficiency, a variance-minimizing filter for nonlinear systems such as particle filters should be applied (van Leeuwen, 2003, 2009). However, the authors leave the investigation of the positive impact of additional correction to EnKF-GEC within the conventional EnKF framework to future work.

2.2.2. Comparisons of forecast skill quality (initial condition quality)

In this subsection, possible reasoning for improvement of EnKF-GEC is hypothesized. Within the perfect model assumption, the quality of forecasts is dependent on that of initial conditions. Therefore, the qualities of initial conditions, which are analysis values at the last analysis time $T+1$, are compared. The initial condition of EnKS or EnKF can be written as follows:

$$X^a_{(T+1|T+1)} = X^f_{(T+1|T)} + P^f_{(T+1|T)}H^T [HP^f_{(T+1|T)}H^T + R]^{-1}\left[y_{T+1} - H\left(X^f_{(T+1|T)}\right)\right].$$

Note that the analysis values of EnKS are the same as that of EnKF at the last analysis time. Meanwhile, the forecast values of EnKF-GEC at time $T+1$ are:

$$X\_GEC^{f}_{(T+1|T+1)} = X^{f}_{(T+1|T)} + [M_{T+1\leftarrow T}][M^{ESV}_{T\leftarrow T+1}]P^{f}_{T+1|T}H^{T} [HP^{f}_{T+1|T}H^{T}+R]^{-1}\left[y_{T+1} - H\left(X^{f}_{(T+1|T)}\right)\right]$$

Note that the equation $P^{f}_{(T+1|T)} = M_{T+1\leftarrow T} P^{a}_{(T|T)} M^{T}_{T+1\leftarrow T}$ is used. Therefore, if the linear empirical operator is the same as the nonlinear operator (e.g. $[M_{T+1\leftarrow T}][M^{ESV}_{T\leftarrow T+1}] = I$), it means that the initial condition of EnKS (or EnKF) is the same as forecast values of EnKF-GEC before performing conventional EnKF at the last analysis time. Therefore, the EnKF procedure at the last analysis time to obtain the initial condition of EnKF-GEC provides the possibility of overweighting of the observations.

However, this overweighting of observations does not always occur, because the additional correction using ESVs does not occur without growing singular modes. The key feature is that the growing singular modes and related additional correction tends to be large when weighting of observations should be larger than that of EnKS due to underestimation of ensemble spread. Figure 2 shows the time series of the ratio of ensemble spread to model error magnitude and magnitude of correction over the Pacific area (130°E–90°W, 10°S–10°N). The variable used is thermocline depth. Note that a negative anomaly of the ratio means that ensemble spread underestimates the model error magnitude. It is clear that there is a negative relationship between them. The correlation coefficient between them is −0.38. It means, in the case of overestimation, that EnKF-GEC can be worse than EnKS due to overweighting of the observations; the magnitude of the additional correction is smaller due to lack of growing SVs, therefore, the solution of EnKF-GEC becomes similar to that of EnKS.

The summary of the above argument is as follows:

(1) The improvement (degradation) of the forecast by using EnKF-GEC occurs when the ensemble spread is underestimated (overestimated).
(2) In the case of ensemble covariance underestimation, EnKF-GEC shows an improvement of analysis quality over EnKS or EnKF, because EnKF-GEC gives more weight to the observations than does EnKS or EnKF.
(3) In the case of ensemble covariance overestimation (or ensemble covariance capturing model error magnitude well), the magnitude of additional corrections using ESVs becomes smaller, because of the relatively smaller magnitude of singular values.

3. Model, and experimental design

3.1. CZ-SPEEDY coupled model

In this study, a coupled model is used to examine the new EnKF algorithm. The oceanic component of the hybrid coupled model is based on an intermediate ocean model similar to the Zebiak–Cane (ZC) model (Zebiak and Cane, 1987). The difference of the present ocean model from the original is that the former uses a new method for parametrization of the subsurface temperature (Kang and Kug, 2000). The atmospheric component of the hybrid model is SPEEDY AGCM (Simplified Parametrizations, primitivE-Equation DYnamics: Molteni, 2003). According to Molteni (2003), the SPEEDY model simulates the general structure of global atmospheric circulation fairly well, and some aspects of the systematic errors are similar to many AGCMs. The resolution of the model is T42L10 (horizontal spectral truncation of 42 wave numbers and 10 vertical levels).

The coupling strategies between pairs of components are as follows. First, the air–sea coupling interval is 10 days. The oceanic (atmospheric) model provides its anomalous SST (anomalous wind stress), and receives anomalous zonal and meridional wind stresses (anomalous SST), whose values are 10-day averaged. Note that the oceanic part of the hybrid coupled model calculates only an anomaly. For details of the coupled model and its performance, see Ham *et al.* (2009).

3.2. Experimental design

Using the CZ-SPEEDY coupled model, data assimilation and seasonal prediction experiments are performed. The experiments are performed in the perfect model context, indicating that one realization of the coupled model is regarded as a true state. To obtain the analysis values (initial conditions for prediction), the EnKF (Evensen, 2003; Leeuwenburgh, 2007) and EnKF-GEC introduced in the previous subsection are employed with 16 ensemble members. The analysis interval is 1 month, and sea-surface temperature (SST) and thermocline depth anomaly are assimilated. In this study, it is simply assumed that observations are located at every grid point. Similarly, the additional correction by EnKF-GEC is also applied to SST and thermocline depth anomaly separately. The prescribed observational errors in the SST and thermocline depth anomaly are 0.35 (°C) and 4 (m), respectively. To reduce the rank deficiency originating from the limited size of the ensemble members, local analysis techniques have been used (Ott *et al.*, 2004). This experiment is identical to that of Ham *et al.* (2009). The assimilation period is 25 years from year 2015 to 2040, then a later 20 years is used to calculate analysis quality.

Using the initial conditions generated by two different EnKF algorithms, a 12-month lead prediction is performed for a 20-year period from year 2021 to 2040. Prediction starts on the 1st of March, June, September and December; thus, the total number of prediction cases is 80. Note that all the 16 ensemble members are used, and then ensemble mean values obtained for these members are used for evaluating the prediction skill. Hereafter, forecasts made using the conventional EnKF are denoted as 'CNTL predictions', and those made with the EnKF-GEC are denoted as 'GEC predictions'.

4. Results

4.1. Analysis results

Figure 3 shows time series of root-mean-square error (RMSE) improvement for the SST and thermocline depth anomaly of EnKF-GEC from conventional EnKF. The detailed procedure to calculate initial conditions in both methods is given in section 2.2.2. Note that the RMSE

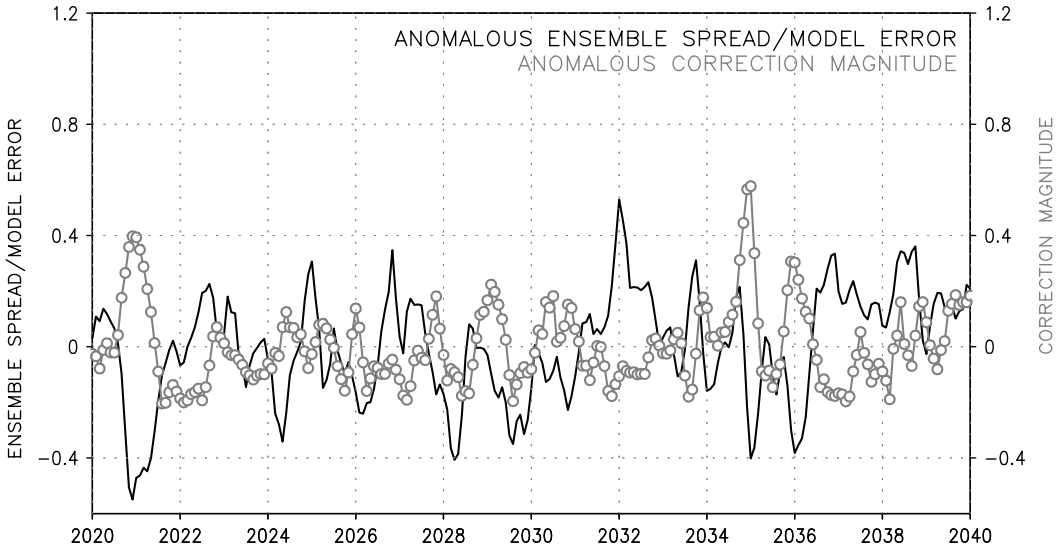

Figure 2. Time series of Pacific area-averaged anomalous magnitude of correction term (grey), and ratio of ensemble spread to model error (black) of thermocline depth. The area-averaged values over 130°E–90°W, 10°S–10°N are shown.

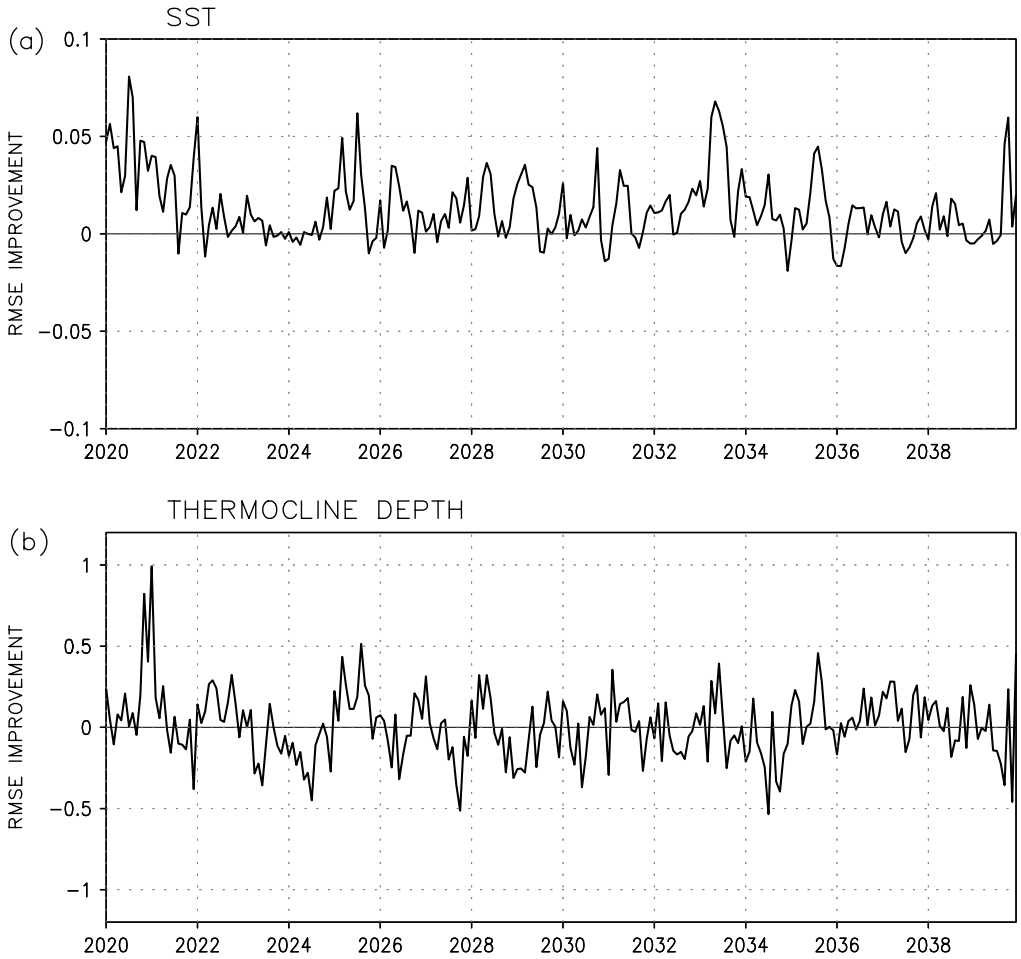

Figure 3. RMS Error (RMSE) improvement for the SST and thermocline depth anomaly. The area-averaged values for 130°E–90°W, 10°S–10°N are shown.

improvement is positive when the analysis error obtained after applying the EnKF-GEC is smaller than that obtained after applying the conventional EnKF. It is clear that the RMSE improvement of the SST anomaly is positive for almost all periods, and hence the new algorithm gives more accurate analysis values. The 20-year-averaged RMSE improvement for the SST anomaly is 0.016, which is approximately 7% of the analysis errors obtained after applying the conventional EnKF. On the other hand, the RMSE improvement of the thermocline depth anomaly is less than that of the SST anomaly. The time-averaged improvement is only around 0.02. Because of the large (small) improvement in the SST (thermocline depth) analysis, the seasonal prediction results obtained using

the EnKF-GEC are superior to those obtained using the conventional EnKF at the early forecast time. This will be discussed later in more detail.

As mentioned before, the distinctive process of the new algorithm is that the magnitude of the correction term is dependent on the underlying instability. This implies that the correction would be more effective when there are robust singular modes during an unstable period. On the other hand, when all singular modes are non-growing (e.g. all singular values are less than 1), correction is not possible, and hence, difference of accuracy in the analysis values obtained using the new algorithm cannot be expected. To investigate the relationship between the magnitude of the correction term and the error growth rate, the area-averaged analysis error for the Pacific region (130°E–90°W, 10°S–10°N) anomalous error growth rate and the magnitude of the correction term for the thermocline depth are calculated. Note that the approximated forecast error ($E(T)$) is denoted as a correction term, and the error growth rate is calculated as follows:

$$\text{Growth rate(T)} = \frac{\int\int \sum_{n=1}^{nrens} [X_n(T+\alpha) - \overline{X(T+\alpha)}]^2}{\int\int \sum_{n=1}^{nrens} [X_n(T) - \overline{X(T)}]^2}.$$

X, nrens, and α denote the state vector (SST), number of ensemble members, and free integration time (1 month), respectively. The correlation coefficient between the two time series is 0.51, indicating that the magnitude of the correction term increases when the ensemble perturbations grow faster than usual. Then, the increased magnitude of the correction terms effectively negates the analysis error in an unstable period, during which the performance of the conventional EnKF generally degrades. This is in agreement with the results obtained by Yang and Kalnay (2010). There is robust improvement in their new 'no-cost' smoother for considering nonlinearity when the growth rate is relatively large. Similarly, the EnKF-GEC developed in this study is effective when the growth rate is large because analysis errors are cancelled out when using unstable empirical singular modes.

Figure 4 shows the maximum covariance analysis (MCA) method between the correction term and analysis errors of the conventional EnKF (also known as singular value decomposition analysis) (Bretherton *et al.*, 1992; Wallace *et al.*, 1992; An *et al.*, 2010), which can be applied for investigating the dominant spatial pattern of the correction terms and analysis error of thermocline depth. If the spatial pattern of the correction term is the same as that of the analysis errors with an opposite sign, it means that the correction method of the new algorithm is working, and negates the analysis error. The first and second MCA modes show similarity between the correction terms and the analysis errors. For easy comparison of the two terms, the sign of the correction term is changed. In the first MCA mode, a negative peak over the central Pacific and a positive peak over the eastern Pacific are observed in both maps; the spatial pattern of the second MCA mode is similar to that of the first mode.

4.2. *Seasonal forecast results*

In order to evaluate the advantage of the EnKF-GEC in seasonal prediction, the forecast skills of the NINO3 SST anomalies for the control (CNTL) and GEC predictions are compared, as shown in Figure 5. The correlation skill and RMSEs are used to evaluate the seasonal forecast performance. The prediction skill observed with EnKF initialization (CNTL predictions) is the same as that with EnKS initialization because the same analysis value is obtained using EnKF and EnKS at the last analysis time. It is clear that the GEC prediction has better forecast skill than the CNTL prediction, particularly during the early forecast lead month. For example, the correlation skill improvement is around 0.1 during a 3-month lead forecast. This large improvement in the forecast skill during the early forecast lead month is because of the accurate SST analysis performed with the EnKF-GEC. The reduction in the SST analysis error is significant, while that in the thermocline analysis error is relatively small; therefore, the prediction skill of the GEC prediction becomes comparable to that of the CNTL prediction as the forecast lead month becomes longer.

Besides the investigation results of the predictability of the NINO index, spatial patterns for the SST correlation skill improvement are shown in Figure 6. The positive sign denotes that the forecast skills of the GEC prediction have improved. The correlation skill improvement differs from region to region. The improvement peak is observed over the equatorial central Pacific at 3-month lead forecast, and then it propagates to the eastern Pacific at 6-month lead forecast. As the forecast lead month becomes longer than 6 months, the improvement is robust over the off-equatorial regions. As shown in Figure 5, during the early forecast lead month, the correlation improves by between 0.05 and 0.1.

One can question as to why the prediction skill improvement differs from region to region. Because of the design and purpose of the new algorithm, the forecast skill improvement would depend on the spatial distribution of the analysis quality achieved using the conventional EnKF. As discussed in the previous subsection, when using the EnKF-GEC, the analysis error is effectively reduced over the region where the conventional EnKF solution is sub-optimal owing to the plural mode. The optimality of the conventional EnKF can be easily evaluated by calculating the correlation skills of the CNTL predictions; this is because there is no model uncertainty in the perfect model context. Hence, the prediction skill is completely dependent on the growth of the initial error.

To investigate this hypothesis, we consider the scatter diagram over the area 190°E–90°W, 10°S–10°N for the 3-month, 6-month and 9-month lead forecast between the correlation skill improvement and correlation skill of the CNTL prediction (Figure 7). It is clear that the correlation improvement is inversely proportional to the correlation skill of the CNTL prediction at the 3-month lead forecast. On the other hand, this inverse relationship is weakened for the longer forecast lead time as the positive impact of EnKF-GEC fades out. For example, there is a weak relationship between correlation skill of the CNTL prediction and correlation improvement at 6-month lead forecasts, and this relationship is barely seen at 9-month lead forecasts. As mentioned earlier, it is related to the fact that improvement of EnKF-GEC is robust in SST, therefore the benefits of accurate analysis are not maintained for long forecast lead months (e.g. longer than 8 months). This means that the inverse relationship does not occur due to plotting methods

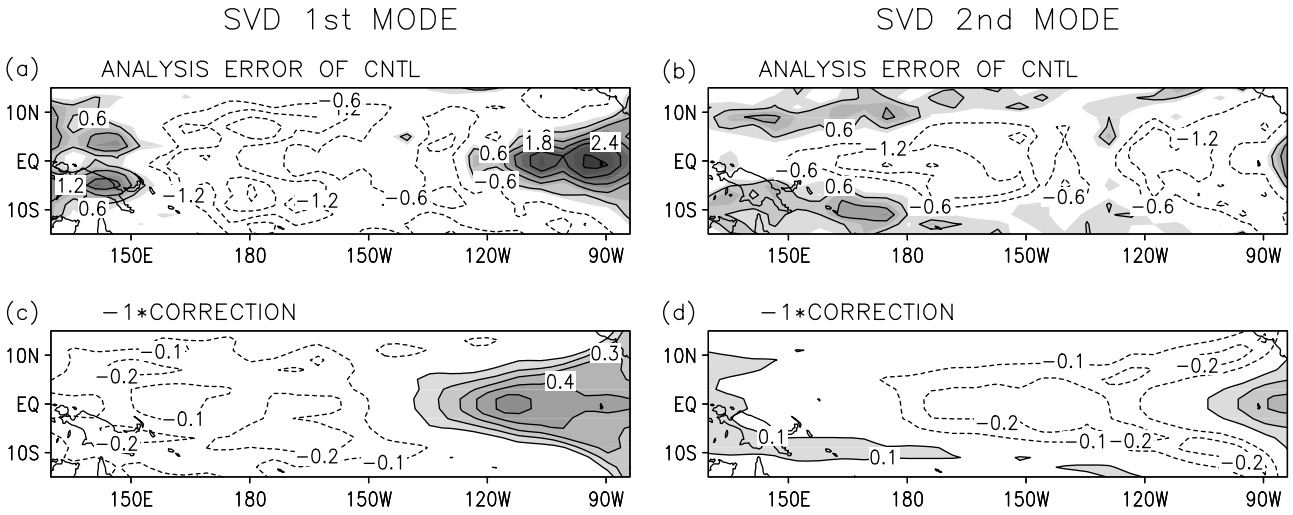

Figure 4. Singular Value Decomposition (SVD) results between the analysis error in the conventional EnKF (upper panels) and the correction term (lower panels) corresponding to the thermocline depth. Note that the negative sign is multiplied to the correction term for easy comparison of the two terms.

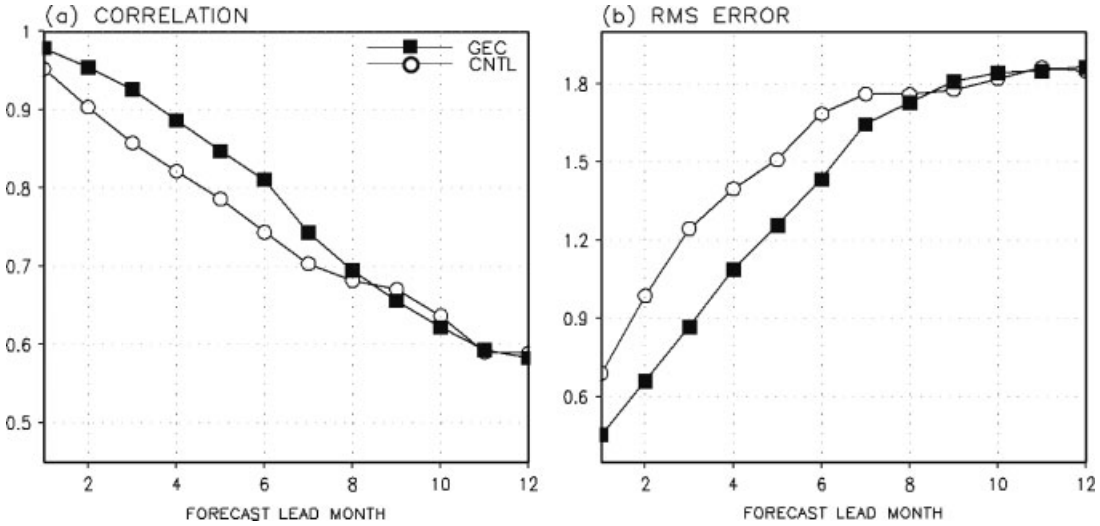

Figure 5. (a) Correlation and (b) RMSEs of NINO3 SST anomalies in the CNTL (filled squares) and GEC prediction (unfilled circles).

or validation methods, but due to positive impacts of EnKF-GEC, whose improvement is robust where the forecast uncertainty is growing fastest. Note that the high (low) correlation coefficient is closely related to the large (small) signal-to-noise ratio, which in turn is related to predictable (unpredictable) directions. This implies that the new EnKF algorithm is more effective for unpredictable regions, where analysis results obtained using the conventional EnKF are sub-optimal.

5. Summary and discussion

In this study, a new EnKF algorithm called EnKF-GEC was developed to assimilate prospective observations using ESVs. For the successful propagation of future information to the target time, the forecast error at a future time is projected onto each final ESV mode, and then the projected magnitude is multiplied with each initial ESV mode. Then, corrected analysis values are obtained by subtracting the forecast errors from the analysis values obtained with the conventional EnKF. The EnKF-GEC differs from the EnKS or 4DEnKF in that the analysis error obtained using the former is significantly reduced until the last analysis time (initial time for prediction). Therefore, seasonal prediction experiments performed with the CZ-SPEEDY coupled model show that the prediction skill whose initial conditions are generated by EnKF-GEC is superior to those generated by the conventional EnKF or EnKS, particularly during the early forecast lead month. In addition, it is shown that the EnKF-GEC algorithm is more effective when used for the unpredictable region, where the analysis values obtained using the conventional EnKF are sub-optimal.

While the performance of the EnKF-GEC is better than that of the conventional EnKF, the computational time in the case of the former is almost twice that in the latter. This is because of the additional time integration step required to calculate the approximated forecast errors at a future time (model forecasts: analysis at future time). Therefore, instead of using the new algorithm, we can increase the number of ensemble members to improve the analysis quality. To make a fair comparison in terms of computational cost, EnKF with 32 ensemble members is additionally performed from year 2015 to 2026, then the time-averaged analysis quality over years 2021–2026 is compared in Table I. For analysis of SST, the RMS error improvement of EnKF-GEC

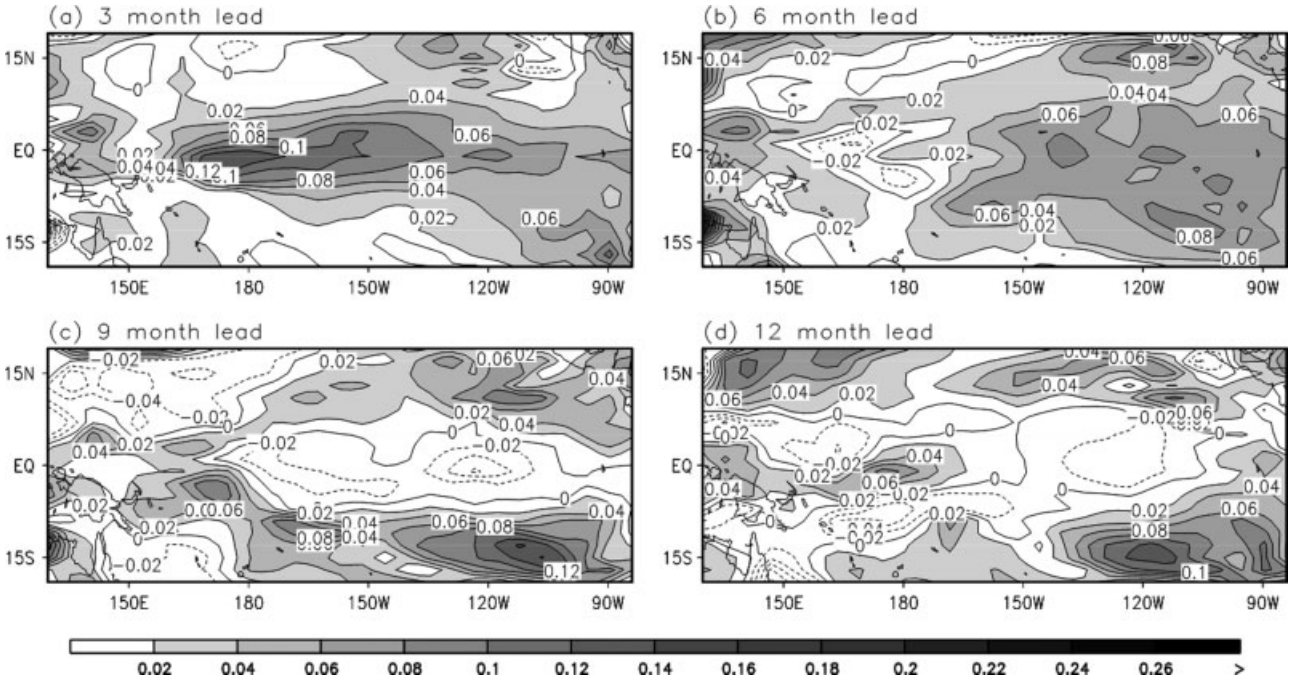

Figure 6. Spatial pattern of correlation skill improvement for the SST anomalies with respect to the forecast lead month. The positive sign denotes that the forecast skills of the GEC prediction are improved. The shading interval is 0.02.

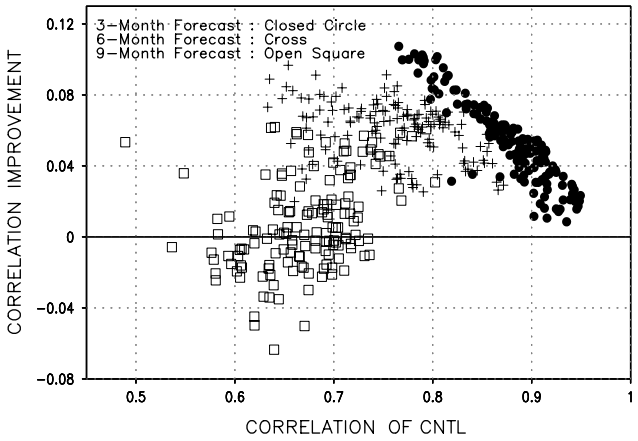

Figure 7. Scatter diagram between correlation improvement (*y*-axis) and correlation skill of CNTL prediction (*x*-axis) for 3-month (closed circle), 6-month (cross), and 9-month (open square) lead forecast over the eastern Pacific region (190°E–90°W, 10°S–10°N).

Table I. Time-averaged RMSE difference of analysis values using EnKF with 32 ensemble members and EnKF-GEC, from that using EnKF with 16 ensemble members. Note that later 6-year-averaged RMSE is used after 10-year assimilation experiments.

RMSE improvement	EnKF-GEC	EnKF with 32 ensemble members
SST (°C)	0.016	0.0029
Thermocline depth (m)	0.039	0.033

and EnKF with 32 ensemble members is 0.016 and 0.0029, respectively, which means that the RMSE improvement of EnKF-GEC is superior to that of EnKF with doubling of ensemble members. On the other hand, the improvement of thermocline depth is similar between doubling ensemble member experiments and the new algorithm. The analysis error reduction is 0.033 and 0.039 for doubling ensemble member experiments and EnKF-GEC, respectively. The reason why EnKF with increased ensemble members does not show the RMSE improvement as expected would be related to the fact that the degrees of freedom of the coupled model are relatively small, due to the simplified physics. For example, five dominant eigenvectors explain over 90% of the total SST variance in the CZ-SPEEDY coupled model.

In addition, the improved performance (analysis error reduction) of the EnKF by increasing ensemble members would be inconsequential if the number of ensemble members of the CNTL prediction were large. This implies that when the number of ensemble members is very large, the analysis quality cannot be increased beyond a certain value by simply increasing the number of ensemble members for the assimilation of the complex GCM. Because the feasible number of ensemble members would be steadily increased by an increase in the computing power, it is difficult to improve the analysis quality of the EnKF by simply increasing the number of ensemble members. Therefore, improvement of the assimilation algorithm is critical. On the other hand, unstable modes, which are calculated empirically by using ensemble members for EnKF, become more accurate as the number of ensemble members increases; therefore, the EnKF-GEC would be more advantageous when the number of ensemble members is increased.

Recently, the 'running in place' algorithm was developed by Kalnay and Yang (2010) to improve the analysis quality at initial spin-up time. In their algorithm, observations were used more than once to extract maximum information when the EnKF was 'cold-started'. They found that it was possible to accelerate the convergence of the EnKF and improve the quality of the EnKF analysis at the early

analysis time. Similarly, in this study, the EnKF-GEC uses the observations more than once when the EnKF solution is considered sub-optimal owing to the significant instability. It is certain that the observations should be used only once when the conventional EnKF system is optimal, because with an optimal EnKF system, usage of observations once can successfully makes the analysis error covariances close to the true error covariance; however, it is noteworthy that both studies mention that modification of this rule is critical for maximum extraction of observational information when the EnKF system is not optimal.

Most new data assimilation algorithms have focused very little on the reduction of the analysis errors at the last analysis time. It is apparent that the improvement of initialization (data assimilation algorithm) is independent of that of the prediction skill. However, this study emphasizes the need for a new data assimilation algorithm beneficial until the last analysis time; this feature is important for prediction accuracy. Therefore, this study lays the foundation for improving the prediction skill within the initialization framework.

Acknowledgements

I.-S. Kang was supported by the Korea Meteorological Administration Research and Development Program under Grant CATER_2006-4206 and the second stage of the Brain Korea 21.

References

An S-I, Ham Y-G, Kug J-S, Timmermann A, Choi J, Kang I-S. 2010. The inverse effect of annual-mean state and annual-cycle changes on ENSO. *J. Climate* **23**: 1095–1110.

Anderson JL, Anderson SL. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **127**: 2741–2758.

Blumenthal MB. 1991. Predictability of a coupled ocean–atmosphere model. *J. Climate* **4**: 766–784.

Bretherton CS, Smith C, Wallace JM. 1992. An intercomparison of methods for finding coupled patterns in climate data. *J. Climate* **5**: 541–560.

Burgers G, van Leeuwen PJ, Evensen G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.* **126**: 1719–1724.

Cohn SE, Sivakumaran NS, Todling R. 1994. A fixed-lag Kalman smoother for retrospective data assimilation. *Mon. Weather Rev.* **122**: 2838–2867.

Evensen G. 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.* **53**: 343–367.

Evensen G, van Leeuwen PJ. 2000. An ensemble Kalman smoother for nonlinear dynamics. *Mon. Weather Rev.* **128**: 1852–1867.

Farrell BF. 1989. Optimal excitation of baroclinic waves. *J. Atmos. Sci.* **46**: 1193–1206.

Fertig EJ, Harlim J, Hunt BR. 2007. A comparative study of 4D-Var and a 4D ensemble Kalman filter: Perfect model simulations with Lorenz-96. *Tellus* **59A**: 96–100.

Gelaro R, Zhu Y. 2009. Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus* **61A**: 179–193.

Ham Y-G, Kug J-S, Kang I-S. 2009. Optimal initial perturbations for El Niño ensemble prediction with ensemble Kalman filter. *Clim. Dyn.* **33**: 959–973.

Harlim J, Hunt BR. 2007. Four-dimensional local ensemble transform Kalman filter: Numerical experiments with a global circulation model. *Tellus* **59A**: 731–748.

Huang X-Y, Mogensen KS, Yang X. 2002. 'First-guess at the appropriate time: The HIRLAM implementation and experiments.' Pp 28–43 in *Proceedings, HIRLAM Workshop on variational data assimilation and remote sensing, Helsinki, Finland*.

Hunt BR, Kalnay E, Kostelich EJ, Ott E, Patil DJ, Sauer T, Szunyogh I, Yorke JA, Zimin AV. 2004. Four-dimensional ensemble Kalman filtering. *Tellus* **56A**: 273–277.

Kalnay E, Yang S-C. 2010. Accelerating the spin-up of Ensemble Kalman Filtering. *Q. J. R. Meteorol. Soc.* **136**: 1644–1651.

Kang I-S, Kug J-S. 2000. An El-Niño prediction system using an intermediate ocean and a statistical atmosphere. *Geophys. Res. Lett.* **27**: 1167–1170.

Kug J-S, Ham Y-G, Kimoto M, Jin F-F, Kang I-S. 2009. New approach for optimal perturbation method in ensemble climate prediction with empirical singular vector. *Clim. Dyn.* **35**: 331–340.

Leeuwenburgh O. 2007. Validation of an EnKF system for OGCM initialization assimilating temperature, salinity, and surface height measurements. *Mon. Weather Rev.* **135**: 125–139.

Lorenc AC. 2003. Modelling of error covariances by 4D-Var data assimilation. *Q. J. R. Meteorol. Soc.* **129**: 3167–3182.

Lorenc AC, Rawlins F. 2005. Why does 4D-Var beat 3D-Var? *Q. J. R. Meteorol. Soc.* **131**: 3247–3257.

Molteni F. 2003. Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Clim. Dyn.* **20**: 175–191.

Moore AM, Kleeman R. 2001. The differences between the optimal perturbations of coupled models of ENSO. *J. Climate* **14**: 138–163.

Ott E, Hunt BR, Szunyogh I, Zimin AV, Kostelich EJ, Corazza M, Kalnay E, Patil DJ, Yorke JA. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus* **56A**: 415–428.

Palmer TN, Buizza R, Molteni F, Chen Y-Q, Corti S. 1994. Singular vectors and the predictability of weather and climate. *Philos. Trans. R. Soc. London* **348**: 459–475.

Penland C. 1996. A stochastic model of IndoPacific sea surface temperature anomalies. *Physica D* **98**: 534–558.

Pitt MK, Shephard N. 1999. Filtering via simulation: Auxiliary particle filters. *J. Amer. Stat. Assoc.* **94**: 590–599.

van Leeuwen PJ. 2003. A variance-minimizing filter for large-scale applications. *Mon. Weather Rev.* **131**: 2071–2084.

van Leeuwen PJ. 2009. Particle filtering in geophysical systems. *Mon. Weather Rev.* **137**: 4089–4114.

Wallace JM, Smith C, Bretherton CS. 1992. Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J. Climate* **5**: 561–576.

Yang S-C, Kalnay E. 2010. Handling nonlinearity and non-Gaussianity in the ensemble Kalman filter. Submitted to *Mon. Weather Rev.*

Zebiak SE, Cane MA. 1987. A model El Niño–Southern Oscillation. *Mon. Weather Rev.* **115**: 2262–2278.

