A Random Subgrouping Scheme for Ensemble Based Filters (sEnSRF)

Yun Liu*[1], Xinyao Rong[1, 2], Zhengyu Liu[1, 3], Shu Wu[1], Shaoqing Zhang[4], Robert Jacob[5]

1 Center for Climate Research and Dept. Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

2 Chinese Academy of meteorological sciences, Beijing 100068, China

3 Lab. of Ocean-Atmos. Studies, Peking University, Beijing 100871, PRC

4 GFDL/NOAA, Princeton University, Princeton, NJ 08542, USA

5 Mathematics and Computer Science Division, Argonne National Laboratory, IL 60439, USA

(to be submitted to MWR)

*Corresponding author
Email: liu6@wisc.edu Tel: 608-261-1459 Fax: 608-263-4190

Abstract

Ensemble based filters can be divided into two categories: stochastic and deterministic. Both types of filters suffer from the problem of generating outliers in the ensembles produced in a nonlinear system. This is especially true for the deterministic filter with a big ensemble as the outliers can persist for a long time, develop into extreme values, and produce large analysis errors.

To address the problem of outliers, a new technique is developed that uses deterministic filter algebra but adds stochastic information into the filter system through random subgrouping. Test results, using the random subgrouping technique on two low-order models (Lorenz-63 and Lorenz-96), show that the new scheme significantly improves performance compared to both stochastic and deterministic filters.

1. Introduction

First introduced by Evensen (1994), the ensemble based filter is emerging as a powerful tool for data assimilation (Evensen 2007). The key element of the filter is to derive the forecast uncertainty from an ensemble of model integrations. The ensemble based filter can be divided into two categories: stochastic and deterministic. The two categories of methods differ mainly in how they update the ensemble after updating the analysis mean. The updated analysis variance is produced by the error variances from both the forecast and observation. A deterministic filter, also called an Ensemble Square Root Filter (EnSRF) (Anderson 2001, 2003, Bishop et al. 2001, Whitaker and Hamill 2002, Tippett et al 2003, Sakov and Oke 2007), transforms the ensemble anomaly to match the variance given by the Kalman Filter (KF) theory (Kalman 1960). In contrast a stochastic filter (EnKF) attempts to match the updated variance from the KF theory by adding perturbations to the observations (Burger 1998, Houtekamer and Mitchell 1998).

The ensemble based filter theory assumes that the background uncertainty and observation noise are Gaussian white and that the ensemble resolves the background uncertainty. These hypotheses introduce two error sources into the analysis: the sampling errors from the limited ensemble size (Whitaker and Hamill 2002, Sacher and Bartello 2008) and the non-Gaussian probability density function (PDF) of the error from a nonlinear system. The sampling error in EnKF appears in both background uncertainty and observational uncertainty. An EnSRF avoids the sampling error introduced by perturbed observations and tends to generate better analyses than an EnKF when applied to a linear model, especially for a small ensemble size (~10-20) (Whitaker and Hamill 2002, Evensen 2003, Anderson 2010).

An EnKF performs better than an EnSRF when applied to a nonlinear system with non-Gaussian PDFs (Lawsen and Hansen 2004, Lei et al 2010, Anderson 2010). A nonlinear system

tends to generate outliers in an ensemble. An outlier is the member that appears to deviate markedly from the general population of ensemble members. An EnKF can mix the outliers by adding random noise to the ensemble members through perturbed observations. Therefore the effect of outliers is relatively weak. An EnSRF tends to have persistent outliers because it has no effective way to restore the outliers to the "right track", thus producing an extreme outlier and large analysis error (Lawsen and Hansen 2004, Anderson 2010). Furthermore, a larger ensemble size is likely to produce more outliers, so an EnSRF applied to a nonlinear system performs worse as the ensemble size increases (Lawsen and Hansen 2004, Mitchell and Houtekamer 2009, Anderson 2010).

Several methods have recently been proposed to improve the performance of the ensemble based filter in eliminating the effect of outliers in non-linear systems. Sakov and Oke (2008) use a random transformation to prevent outliers in the ensemble transform filter, while Anderson (2010) uses a rank histogram filter to eliminate outliers. In this study a new filter scheme, called the random subgrouping Ensemble based filter (sEnSRF), is proposed to eliminate the distortion effect of outliers and therefore to improve the filter performance in a nonlinear system. sEnSRF randomly divides the entire ensemble into subgroups of equal size and updates each subgroup independently using EnSRF. In comparison with EnSRF and EnKF, sEnSRF significantly improves the filter analysis in two tested nonlinear systems. In section 2, the algorithm of the ensemble based filter and sEnSRF are briefly described. We will demonstrate the performance of sEnSRF in simple concept models in section 3. A summary will be given in section 4.

2 The Random Subgroup Ensemble Based Filters

Data assimilation statistically merges model forecasts with observations to generate analyses with a reduced error. It proceeds by analysis cycles that consist of two steps: a forecast step and an analysis step. We define $x = x(t)$ to be a n-dimensional column vector for the model state at time t and y_t to be the observation at time t. In an analysis step we solve the conditional probability distribution density function (PDF) $P(x(t) \mid Y_t)$ of model states at time t, where $Y_t = [y_1, y_2, y_3, \ldots, y_{t-1}, y_t]$. Based on Bayes' rule, the conditional PDF can be written as

$$P(x(t) \mid Y_t) = P(y_t \mid x(t)) P(x(t) \mid Y_{t-1}) / P(y_t \mid Y_{t-1}) \quad (1)$$

In a forecast step we integrate the forecast model from x(t) and $P(x(t) \mid Y_t)$ to obtain x(t+1) and $P(x(t+1) \mid Y_t)$, where t+1 is the next observation time.

Kalman Filter

The kalman filter (KF) achieves an optimal estimation of $x(t)$ for either a condition of minimum variance or a maximum likelihood based on the assumption of the error distributions that follow Gaussian PDFs. The PDFs are represented by the mean and covariance,

$$\begin{aligned} &P(x(t) \mid Y_t) \sim N(x^a, P^a) \\ &P(x(t) \mid Y_{t-1}) \sim N(x^f, P^f) \qquad (2) \\ &P(y_t \mid x(t)) \sim N(y^o, R) \end{aligned}$$

where x^a, x^f, y^o are the analyses, forecasts and observations (denoted by a, f, o separately); P^a, P^f, R are the corresponding covariance and N denotes the normal distribution.

The KF updates the analysis (x^a) and covariance (P^a) as

$$x^a = x^f + K(y^o - Hx^f) \quad (3)$$

$$P^a = (1 - KH)P^f \quad (4)$$

where $K = P^f H^T (HP^f H^T + R)^{-1}$ is the Kalman gain. H is the linearized mapping function from

state variables space to observation space, y=Hx and H^T is the H transpose. The observational uncertainty R is determined by the observations itself. The forecast uncertainty P^f is advanced from the previous P^a by using the Fokker–Planck Kolmogorov equation (Jazwinski 1970), which is difficult to apply in high dimension sytems like weather and climate models.

EnKF

Different from KF, the ensemble based method uses a Monte Carlo scheme to generate an ensemble $x_i^a(t)$ to sample $P(x(t)|Y_t)$. It integrates the ensemble to determine $x_i^f(t+1)$ and $P(x(t+1)|Y_t)$, where the ensemble index i = 1, 2,…, N and N is the ensemble size.

A stochastic filter (EnKF) treats observations as random variables that are perturbed to sample the observational uncertainty. The resulting analysis variance will match what is derived theoretically with KF (sea eqn. 4, Burgers et al 1998, Houtekamer and Mitchell 1998). To ensure no extra noise is added by the perturbations, the ensemble mean perturbation is set equal to zero (Pham 2001; Mitchell et al. 2002, Evensen 2007, Mitchell and Houtekamer 2009).

$$X_i^a = X_i^f + K(y_i - HX_i^f) \quad (5)$$

EnSRF

The Deterministic filter (EnSRF) transforms the forecast ensemble to match the analysis and its uncertainty in eqn. (3) and (4). This paper will show the results from one particular EnSRF, the Ensemble adjustment filter (EAKF, Anderson 2001, 2003). EAKF updates the ensemble in two steps (Anderson 2003). First it derives the analysis ensemble mean and variance and then computes the ensemble increment to match the analysis error in the observation space. In the second step, the ensemble increment is distributed over relevant state variables using a least square fitting.

sEnSRF

In this study, a new filter scheme, called random subgrouping EnSRF (sEnSRF), has been developed to eliminate the effect of persistent outliers that often occur in a regular EnSRF. This scheme is the same as a regular EnSRF except that at each analysis step, the entire ensemble is randomly divided into sub-ensembles of equal size. In sEnSRF, all the sub-ensembles are updated independently using the same observations, but with their own forecast covariance. Each sub-ensemble will have a different combination of ensemble members at different analysis time. The analyses and forecasts are constructed using the entire ensemble.

The objective of our sub-grouping procedure differs from that of Houtekamer and Mitchell (1998), where they use a sub-grouping process to reduce the negative bias in the analysis error variance. Here, the random subgrouping is used to introduce random information into filter system. A sEnSRF is performed in 4 steps, as shown schematically in Figure 1:

(1) the model ensemble is integrated forward to the next observation time;

(2) the N-member ensemble is randomly divided into n sub-ensembles of equal N/n size;

(3) each sub-ensemble is updated independently using EnSRF with the same observations;

(4) steps 1 to 3 are repeated until the all the observations are used.

In summary, the sEnSRF shares the same algebra as EnSRF, but adds some stochastic variability through random subgrouping. It is still a square root method but no longer deterministic. We will show that sEnSRF performs better than EnKF and EnSRF in a strongly nonlinear system because it avoids the effect of outliers that often occurs in EnSRF and the sample error that is introduced by perturbing observations in EnKF.

3. sEnSRF in the Lorenz model

As a test, we apply the random grouping EAKF to the Lorenz63 model system (Appendix A). A random subgrouping EAKF with *n* subgroups will be denoted as $sEAKF_n$. Our $sEAKF_n$ results will be compared with the results from EnKF and EAKF applied on the same Lorenz system using identical observations and initial conditions. The details of experimental design are shown in appendix A.

A) Big ensemble experiments

For the test conducted here we us an ensemble size of 80 and a subgroup size of 16. Due to the strong nonlinear nature of the Lorenz63 model, outliers are produced in the 80-member ensemble for the EnKF and EAKF simulations but not in $sEAKF_{16}$ simulation with 16 subgroups (Fig. 2). In EAKF (Fig.2b), an extreme outlier persists for more than 200 analysis steps with the deviation far from the other ensemble members. This derivation occurs because of the lack of a mechanism in EAKF that prevents drifting from the ensemble mean. An EAKF tends to retain high-order moments through the assimilation process (Anderson 2001). While this feature may be a good choice for a filter problem, it leads to the outliers persistent during the EAKF assimilation process. The persistence of outliers can produce big separation from most other members leading to extreme outliers, as shown in Fig. 2b.

The outliers are less extreme in EnKF (Fig.2a) than in EAKF (Fig.2b), because random perturbations in the observations play the role of adding random noise to outliers in EnKF. This works to restore an outlier toward the ensemble mean after a few cycles of analysis.

The random subgrouping for $sEAKF_{16}$ plays a similar role in preventing persistent outliers that was accomplished using perturbed observations for EnKF (Fig.2c). An ensemble member that is an outlier in one subgroup combination may in the next step be a regular member

for another subgroup combination. As a result, an outlier at one analysis step can be eliminated by the random subgrouping at the next analysis step. Furthermore, the $sEAKF_{16}$ is less likely to produce an outlier than EnKF and EAKF because each sub-ensemble has a much smaller sample size than the full ensemble and a small sample size produces less outliers.

A good measure for the presence of outliers is the ensemble kurtosis

$$Kur = \frac{\sum_i (x_i - \bar{x})^4}{(\sum_i (x_i - \bar{x})^2)^2} \qquad (6)$$

where $\bar{x}$ is ensemble mean. The time-mean kurtosis of y is ~20 for the EAKF with an 80-member ensemble. This is much larger than the kurtosis value of 3 for the Gaussian distribution and the system kurtosis value 2.5 for the y-variable of Lorenz63 system derived from 10^4 time units control integration. The time-mean kurtosis of y is ~4 for EnKF for the 80-member ensemble, which is much less than EAKF and a little greater than Gaussian. In comparison, the kurtosis of y in $SEAKF_{16}$ is ~2.5, which matches the system kurtosis and indicates that few spurious outliers are generated by this assimilation scheme. Therefore, $sEAKF_{16}$ has a reduced outlier problem relative to EnKF and, especially, EAKF. Only the $sEAKF_{16}$ preserves the system 4th-order moment.

The ensemble spread represents the uncertainty (error) of the ensemble mean in an ensemble-based filter. We expect a consistent ensemble spread and RMSE during assimilation. The $sEAKF_{16}$ and EAKF simulations do show statistically consistency between RMSE and ensemble spread (Fig.3a). Because persistent outliers distort PDFs, EAKF produces a big difference between RMSE and ensemble spread. Some EAKF experiments generate very big analysis error (with the RMSE >1) but very small ensemble spread (~ 0.7), so the average RMSE is much smaller than the ensemble spread (Fig.3a).

The $sEAKF_{16}$ generates the smallest analysis error among the three filter schemes. $sEAKF_{16}$ reduces the root mean square error (RMSE) of the analysis to 0.58, which is significantly smaller (with 99% confidence) than the RMSE for EAKF (0.75) and EnKF (0.62) (Fig.3b). Starting from the idential first guess (initial conditions) and using idential observations, 99% (90%) of $sEAKF_{16}$ experiments produce smaller RMSE than corresponding EAKF (EnKF) experiments (Fig.3b).

B) Small ensemble experiments

The $sEAKF_n$ scheme also improves filter performance on the Lorenz63 system for smaller ensemble size (<80). EAKF simulations with a 20-member ensemble, which have an average ensemble kurtosis of about 3.6, still suffer from the effect of outliers. $sEAKF_4$ (4 subgroups) simulations with a 20-member ensemble produce smaller RMSE (0.59 vs. 0.64) (with 99% confidence) than corresponding EAKF simulations (Fig.4b). Because of the small ensemble size, the effect of sampling error from the perturbed observations in EnKF simulations exceeds the outlier effect in the EAKF simulations. The EnKF simulations perform poorest among the three schemes (Fig.4). The small ensemble size also leads to an ensemble spread that is smaller than RMSE, especially for EnKF (Fig 4a). For an idential first guess (initial conditions) and observations, more than 80% of $sEAKF_4$ experiments produce smaller RMSE than corresponding EAKF and EnKF experiments (Fig.4b).

For both big ensemble and small ensemble experiments, $sEAKF_n$ significantly improves the data assimilation quality in Lorenz63 system compared with EnKF and EAKF.

C) Optimal size of the sub-group

The number of subgroups (n) is a free parameter in $sEAKF_n$ and it is desirable to know what is the optimal subgroup number. The random subgrouping scheme eliminates the effect of

outliers by introducing randomness into the filter system. In fulfilling this goal, $sEAKF_n$ requires sufficient stochastic degrees of freedoms and small enough sub-ensemble sizes. If the subgroup is too small, the effect of outliers will still exist though significantly reduced. For example, the ensemble kurtosis of $sEAKF_2$ (2 subgroups) is ~4.2 for the 80-member ensemble experiments and ~7.7 for 160-member ensemble experiments (Fig. 5a). However, the number of subgroups is bounded by the ensemble size with biggest subgroup being N/2 for a N-member ensemble.

The sample size for the sub-ensembles cannot be too small. When the maxim subgroup is used and each subgroup has only 2 members, the ensemble kurtosis converges to the kurtosis of the Gaussian distribution 3, instead of the Lorenz63 system kurtosis of 2.5 (Fig.5a). This occurs because each subgroup has only two members, making it impossible to resolve a PDF and the higher-order moments beyond that of Gaussian. (The convergence to the Gaussian value results from averaging a large number of subgroups). As a result, both the 2rd-order and 4th-order moments are distorted.

A larger ensemble size increases the likelihood for outliers and therefore requires more subgroups to eliminate them. A larger ensemble size also accommodates more subgroups. As a result, for different sizes of ensembles, the minimum kurtosis seems to be achieved at a size of sub-ensemble independent of full ensemble size (Fig.5a). For ensemble sizes of 20, 40, 80 and 160, the minimum kurtosis converges to the system kurtosis of about 2.5 for a sub-ensemble size of 5 (Fig. 5a). This is the optimal subgrouping for conserving the 4th-order moment and removing the effect of outliers.

A smaller sampling kurtosis reduces the outliers and therefore improves the filter performance (see RMSE – Fig5b). Thus, the filter performance largely follows the sample kurtosis for our system. Smaller RMSE occurs for smaller sample kurtosis. For the 80-member

ensemble size, $sEAKF_{16}$ gives the smallest analysis RMSE, which is reduced by ~20% from that of the standard EAKF. While sub-sample (N/n=5) is the optimal sub-sample size, it is worth noting that even for small subgroup size of 2, $sEAKF_2$ can significantly decrease ensemble kurtosis and analysis RMSE compared to EAKF (Fig5).

D) Sampling error for subgrouping

Although successfully removing the outliers and maintaining the higher-order moments, random subgrouping scheme introduces sampling errors into the filter simulation. Since each sub-ensemble has a much smaller sample size than the full ensemble, the background uncertainty estimated from an individual sub-ensemble has greater sampling error than that estimated from the full ensemble. However, the PDFs constructed from all the sub-ensembles automatically represent the uncertainty of the forecast error PDFs. This leads to a compensation of the sampling error generated by the subgrouping. Therefore the net increase in sampling error is lessened.

A Monte Carlo method is used to evaluate the sampling error caused by subgrouping (Figure 5). 100 samples are divided into n groups of 100/n samples to resolve a Gaussian PDF. The variance is calculated independently for each subgroup and then averaged from all subgroups. The sampling error of variance from the limited sample size (here is 100) is represented by the standard deviation of variances from 100,000 Monte Carlo experiments.

The uncertainty of the expected variance resulting from sampling error for a Gaussian PDF constructed from 100 samples increases from 14% to 16% of the total variance as the subgroup number increases from 1 to 20. This represents only a 2% increase in the total variance, which is negligible compared with the effect of outliers in EAKF. As a result, the $sEAKF_n$ performs significantly better than EAKF.

E) Subgrouping in EnKF

EnKF also benefits from random subgrouping for big ensemble experiments because there are weak effects from outliers in EnKF simulations. The kurtosis of EnKF is ~4 for the 80-member ensemble experiments and ~5 for 160-member ensemble experiments (Fig5). To test the subgrouping scheme, we apply an 8 member random subgrouping to the 160-members EnKF simulation using the Lorenz63 model. Results show RMSE on the random subgrouping EnKF decreases ~2% and the kurtosis decreases to ~3 compared to EnKF (figure not show).

F) sEnSRF in Lorenz96 model – model independence

The effectiveness of $sEAKF_n$ for systems of different levels of chaos is also examined by using different forcing (F) in the Lorenz96 model with 200 dimensions (details in Appendix B). When the system is near neutral (F=2), the outlier effect is minor for an 80-member ensemble simulation. In this case, the EAKF performs better than EnKF. The RMSEs for EAKF, $sEAKF_2$ and $sEAKF_4$ are comparable, implying insensitivity to the subgrouping scheme. As the system becomes chaotic (F = 5, 8, 10), the sEAKFn simulations with an 80-member ensemble perform significantly better (with 99% confidence) than the corresponding EAKF and EnKF simulations. The RMSE from sEAKFn with optimal subgroups (4 or 8) is reduced by 7~9% from that of the standard EAKF and is reduced by 4~5% from that of the EnKF (Table 1).

In summery, the improvement in results of $sEAKF_n$ over EAKF and EnKF for different models (Lorenz63 and Lorenz96) suggests that our filter scheme is a robust method that should be tested on complex weather and climate models.

4. Summary

The persistent outlier problem arising from non-Gaussian PDFs is a challenge for an ensemble based filter, especially for EnSRF using large ensemble size (Lawsen and Hansen 2004, Anderson 2010, Lei et al 2010). Using a random subgrouping scheme with EnSRF, our sEnSRF scheme reduces the outlier problem and significantly improves the data assimilation quality in nonlinear systems. The sEnSRF uses the same algebra as EnSRF, but is no longer a deterministic filter because of the random grouping. It avoids sampling errors introduced by perturbations in the observations, eliminates the effect of outliers, and retains the higher-order moments.

The random subgrouping eliminates outliers in two ways. First, the smaller size of each sub-ensemble leaves less chance to produce outliers compared with the full ensemble. Second, the random subgrouping eliminates the existing outliers by introducing randomness into the filter system.

As a test, a sEnSRF ($sEAKF_n$) is applied to two simple models: the Lorenz63 and Lorenz96 model. The random subgrouping significantly improves the filter analysis relative to both the stochastic filter EnKF and the deterministic filter EnSRF.

sEnSRF has the advantage of being simple as well as practical. It can be easily applied to a higher dimension system. In particularly, it is so effective in highly chaotic systems (like Lorenz63 and Lorenz96) that tend to generate more extreme outliers. Because of this result, we propose testing sEnSRF in data assimilation for complex weather and climate models.

Acknowledgements. We would like to thank Dr. R. Gallimore for his help on writing and English. We would also like to thank Dr. T. Hamill for his comments on an earlier version of the manuscript. This work is supported by NSF and 2012CB955201 and was also supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

Appendix A Lorenz model

The Lorenz model (Lorenz 1963) describes one of the most famous nonlinear dynamical systems

$$\dot{x} = -\sigma(y - x)$$

$$\dot{y} = \beta x - y - xz$$

$$\dot{z} = xy - cz$$

The parameter β is the ratio of the Rayleigh number divided by the critical Rayleigh number. The parameter σ is the Prandtl number. The third parameter c is related to the horizontal wave number of the system. By choosing typical values of the parameters (β = 28, σ=10, c=8/3), the evolution of the state vector (x, y, z) follows the well-known Lorenz attractor.

The model is integrated using a 4-th order Runge-Kutta method with a time resolution of *dt=0.01* (~1 hours if we treat one time unit as 4 days). We first generate the "truth" in a long control simulation of 10^4 time units.

For each set of experiment, we first generate the "observations" by adding onto the "truth" random errors with a standard derivation (2, 2, 2). The initial first guess is then randomly picked from the "observations". The initial ensemble is built based on the initial first guess and its uncertainty (error scale). Using the initial ensemble and "observation", the model is simulated for 50 time units (500 analysis steps) by using different data assimilation schemes ($SEAKF_n$, EAKF, EnKF) with an observation time interval use 0.1. Following a 100 analysis steps spinup, the results from all analysis steps are used to calculate the RMSE, kurtosis, ensemble spread for the experiment.

500 experiments are performed for each experiment setting. For a fare comparison we did not

use any inflation scheme in all the experiments.

Appendix B

The Lorenz96 model is a latitude circle model first proposed by Lorenz (1996) to study fundamental issues regarding the forecasting of spatially extended chaotic systems such as the atmosphere. It has N state variables governed by equation

$$\dot{X}_i = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F$$

where i = 1,…,N are the cyclic indices. We use N = 200 for our simulations. The model is integrated using a 4-th order Runge-Kutta method with a time resolution of dt=0.005. To investigate the filter performance under different conditions, we assign values (2, 5, 8, 10) for the forcing term F which defines the system shifting from near neutral (2) to strong chaos (5, 8, 10).

The experiment design is similar to Appendix A but all the simulations use an ensemble size 80. The observations cover all the grids with a frequency 0.1 (20 time steps) and the observation error standard derivation 2. The data assimilation length for each experiment is 50 time units (500 analysis steps) and first 10 time units (100 analysis steps) are treated as spinup. The covariance localization with an influence radius of 11 is used in all the simulation (Mitchell et al. 2001). The RMSE are calculate from all analysis steps after spinup and averaged for all the grids. 500 experiments are performed using each scheme and no inflation was used in the experiments.

Reference

Anderson, J., 2001: An ensemble adjustment kalman filter for data assimilation. Monthly Weather Review, 129, 2884–2903.

Anderson, J. L., 2003: A local least squares framework for ensemble filtering. Monthly Weather Review, 131, 634–642.

Anderson, J. L., 2010: A non-Gaussian ensemble filter update for data assimilation. Monthly Weather Review, 138,4186–4198.

Bishop, C. H., B. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transformation kalman filter. part i: theoretical aspects. Monthly Weather Review, 129, 420–436

Burgers, G., P. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. Mon. Wea. Rev., 126, 1719–1724.

Evensen, G., 1994: Sequential data assimilation with a non-linear quasi-geostrophic model using monte carlo methods to forecast error statistics. J. Geophys. Res., 99(C5), 10 143–10 162.

Evensen, G., 2003: The ensemble kalman filter: theoretical formulation and practical implementation. Ocean Dynamics, 53, 343–367

Evensen, G., 2007: Data assimilation: the ensemble Kalman filter. Springer.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data assimilation using an ensemble kalman filter technique. Monthly Weather Review, 126, 796–811.

Jazwinski, A.H. 1970: Stochastic processes and filtering theory. Academic Press, 376 pp.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82, 34-45.

Lawson, G. W. and J. A. Hansen, 2004: Implications of stochastic and deterministic filtersas ensemble-based data assimilation methods in varying regimes of error growth. Monthly Weather Review, 132, 1966–1981.

Lei, J. and P. Bickel, 2010: Comparison of Ensemble Kalman Filters under Non-Gaussianity. Monthly Weather Review, 138, 1293–1304.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20, 130–141.

Lorenz, E. N., 1996: Predicbility: a problem partly solved. In: Proceedings' of the ECMWF seminar on predibility, vol1, ECMWF, Reading, United Kingdom, 1-18.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Mon. Wea. Rev., 129, 2776–2790

Mitchell, H. L. and P. L. Houtekamer, 2009: Ensemble Kalman Filter Configurations and Their Performance with the Logistic Map. Mon. Wea. Rev.,137, 4325-4343.

Mitchell, H. L. and P. L. Houtekamer and G. Pellerin, 2002: Ensemble size, balance, and model-error representation in an ensemble Kalman filter.Mon. Wea. Rev., 130, 2791–2808.

Pham, D. T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon. Wea. Rev., 129, 1194–1207.

Sacher, W. and P. Bartello, 2008: Sampling errors in ensemble Kalman filtering. Part i: theory. Monthly Weather Review, 136, 3035–3049.

Sakov, P. and P. R. Oke, 2007: Implications of the form of the ensemble transformation in the ensemble square root filters. Monthly Weather Review, 136, 1042–1053.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. Monthly Weather Review, 131, 1485–1490.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbedobservations. Monthly Weather Review, 130, 1913–1924.

Table

Table 1 The mean analysis RMSE for $sEAKF_n$, EAKF and EnKF derived using the Lorenz96 system from 500 experiments with 80-member ensembles. The first column denotes the forcing term in the system.

	$sEAKF_8$	$sEAKF_4$	$sEAKF_2$	EAKF	EnKF
F=2	0.032	0.030	0.030	0.030	0.033
F=5	0.336	0.338	0.345	0.368	0.349
F=8	0.657	0.656	0.666	0.705	0.686
F=10	0.759	0.755	0.767	0.810	0.793

Figure Caption List

Figure 1 Flow chart of the sEnSRF data assimilation system for the case of 25-member ensemble and 5 subgroups. Each small square demotes an ensemble member and each row denotes a subgroup. Each initial subgroup (left hand side) has 5 samples identified with the same color. And later the procedure randomly mixes the members into new subgroups identified with a mix of colored members.

Figure 2 Initial error evolution of 80 ensemble members (black lines) and the ensemble means (red lines) for variable y in Lorenz system. (a) is for EnKF simulation; (b) is for EAKF simulation and (c) is for $SEAKF_{16}$ simulation. (The figure follows figure 3 in Anderson 2010).

Figure 3 The scatter diagram for 500 $sEAKF_{16}$ experiments with an 80-member ensemble and the corresponding EnKF (EAKF) simulations.

(a) The scatter diagram of analysis RMSE and ensemble spread of variable y for different filter schemes. The red dots are for EAKF simulation, green circles are for EnKF and blue stars are for $sEAKF_{16}$. The squares represent the average of a total of 500 experiments. A point on the black line denotes the situation when the RMSE is equal to the ensemble spread.

(b) The scatter diagram of analysis RMSE for EnKF vs $sEAKF_{16}$ (red dot) and EAKF vs $sEAKF_{16}$ (green star). The x-axis is the analysis RMSE for EAKF and EnKF simulation and the y-axis is the RMSE for $sEAKF_{16}$. The squares represent the average of a total 500 experiments. A point on the black line denotes when the two schemes give the same RMSE.

Figure 4 The same as figure3 except for $sEAKF_4$ experiments with a 20-member ensemble and the corresponding EnKF (EAKF) simulations.

Figure 5 The ensemble kurtosis and analysis RMSE for different ensemble size averaged from 500 experiments.

The x-axis represents the sample sizes for subgroups of $sEAKF_n$ (EnKF, EAKF). The blue dot lines are for 20-member ensembles; green plus lines are for 40-member ensembles; red circle lines are for 80-member ensembles; and cyan star lines are for 160-member ensembles. The

squares represent the results from EnKF simulations and the diamonds represent the average results from EAKF simulations. The two black dish lines on upper panel are the kurtosis for Lorenz63 system (2.5) and for a Gaussian white noise distribution (3).

The kurtosis (RMSE) from EAKF for an ensemble size of 80 and 160 are 14.5 (0.75) and 56.8 (1.05) that are too large to be shown on the plots.

Figure 6 The variance uncertainty generated by limited sampling and different subgrouping for a Gaussian PDF with variance 1. The variance uncertainty derives from 100,000 Monte Carlo realizations

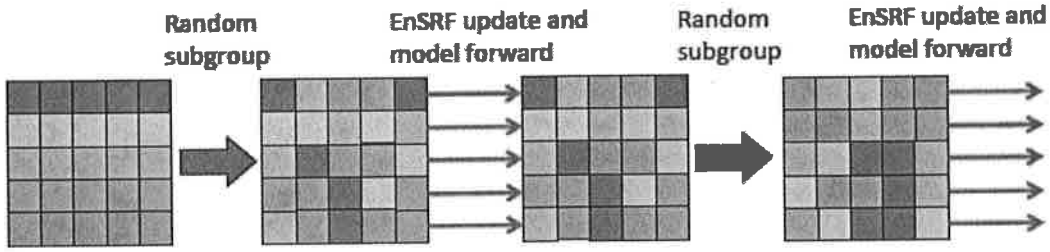

Figure 1 Flow chart of the sEnSRF data assimilation system for the case of 25-member ensemble and 5 subgroups. Each small square demotes an ensemble member and each row denotes a subgroup. Each initial subgroup (left hand side) has 5 samples identified with the

same color. And later the procedure randomly mixes the members into new subgroups identified with a mix of colored members..

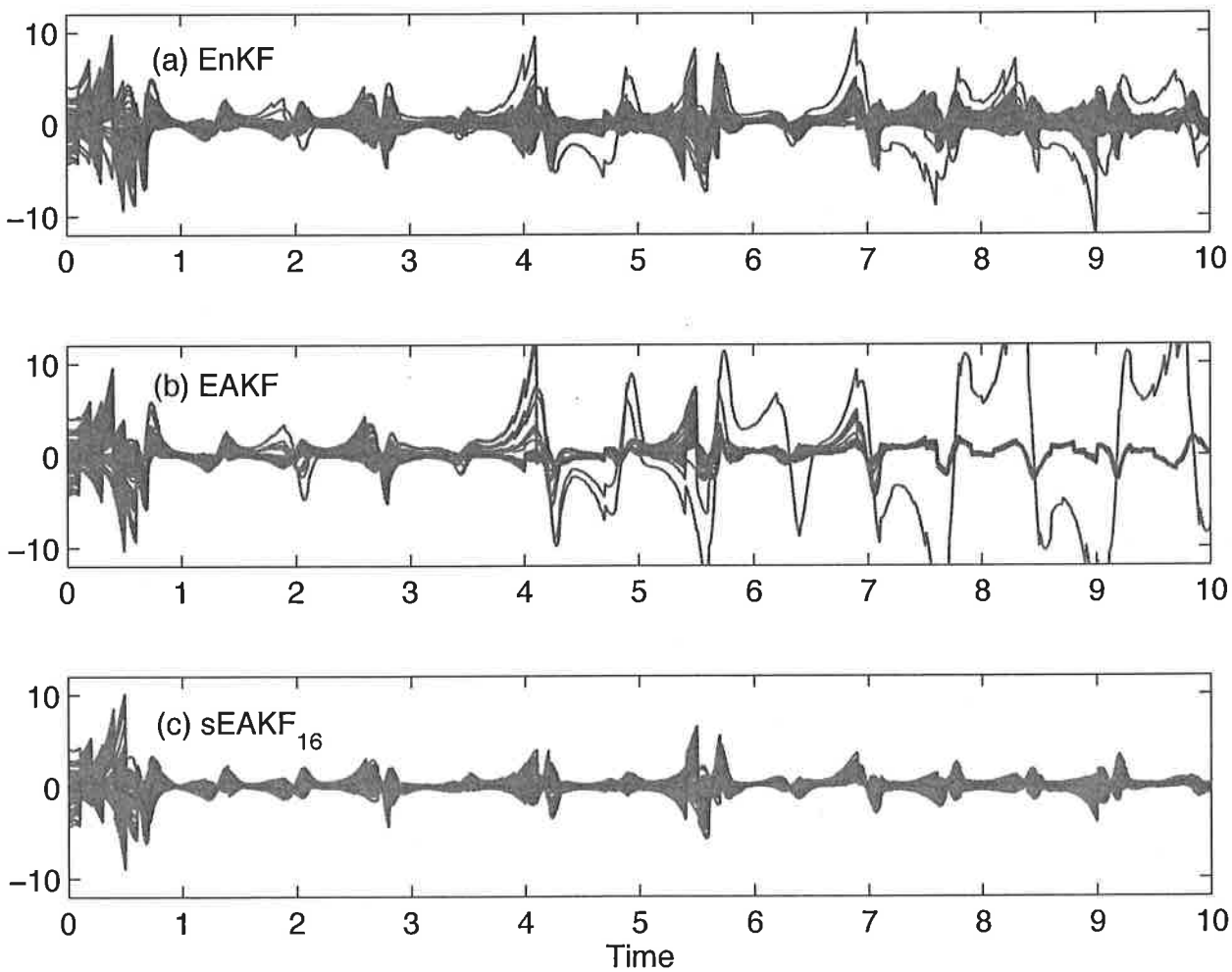

Figure 2 Initial error evolution of 80 ensemble members (black lines) and the ensemble means (red lines) for variable y in Lorenz system. (a) is for EnKF simulation; (b) is for

EAKF simulation and (c) is for $SEAKF_{16}$ simulation. (The figure follows figure 3 in Anderson 2010).

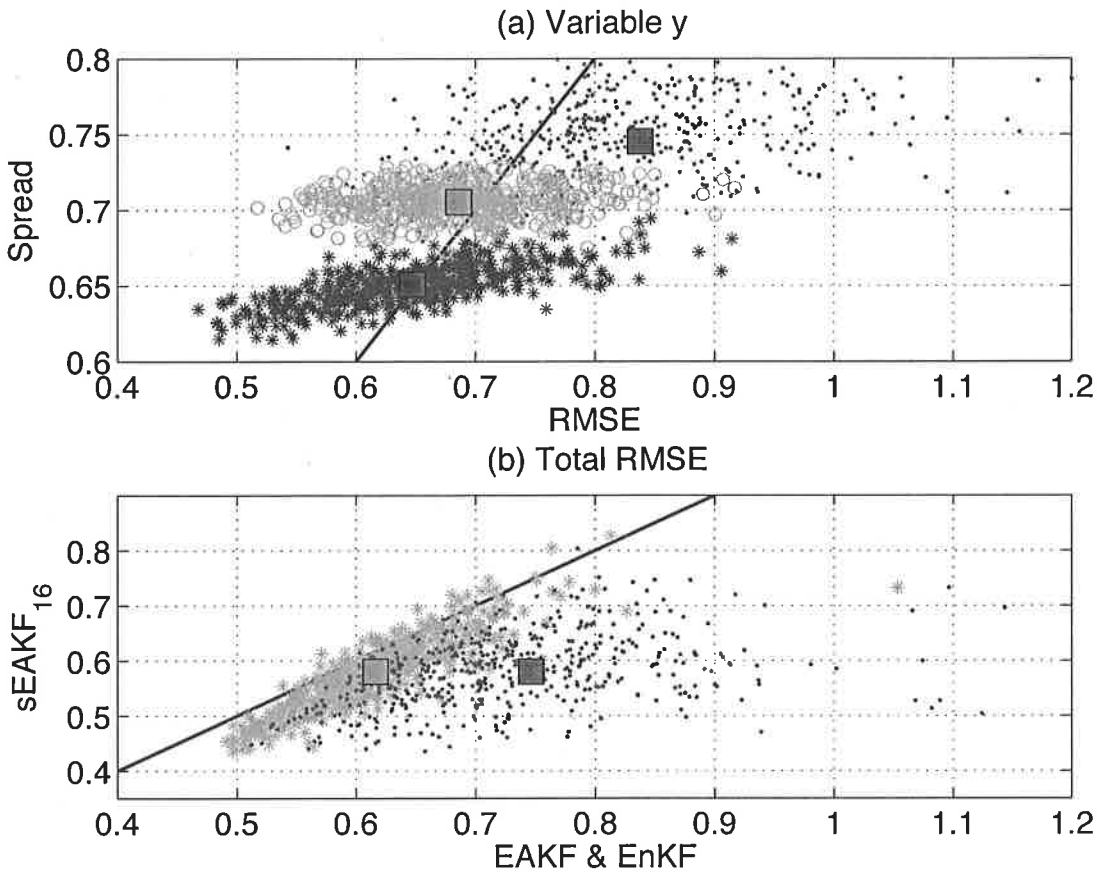

Figure 3 The scatter diagram for 500 $sEAKF_{16}$ experiments with an 80-member ensemble and the corresponding EnKF (EAKF) simulations.

(a) The scatter diagram of analysis RMSE and ensemble spread of variable y for different filter schemes. The red dots are for EAKF simulation, green circles are for EnKF and blue stars are for $sEAKF_{16}$. The squares represent the average of a total of 500 experiments. A

point on the black line denotes the situation when the RMSE is equal to the ensemble spread.

(b) The scatter diagram of analysis RMSE for EnKF vs $sEAKF_{16}$ (red dot) and EAKF vs $sEAKF_{16}$ (green star). The x-axis is the analysis RMSE for EAKF and EnKF simulation and the y-axis is the RMSE for $sEAKF_{16}$. The squares represent the average of a total 500 experiments. A point on the black line denotes when the two schemes give the same RMSE.

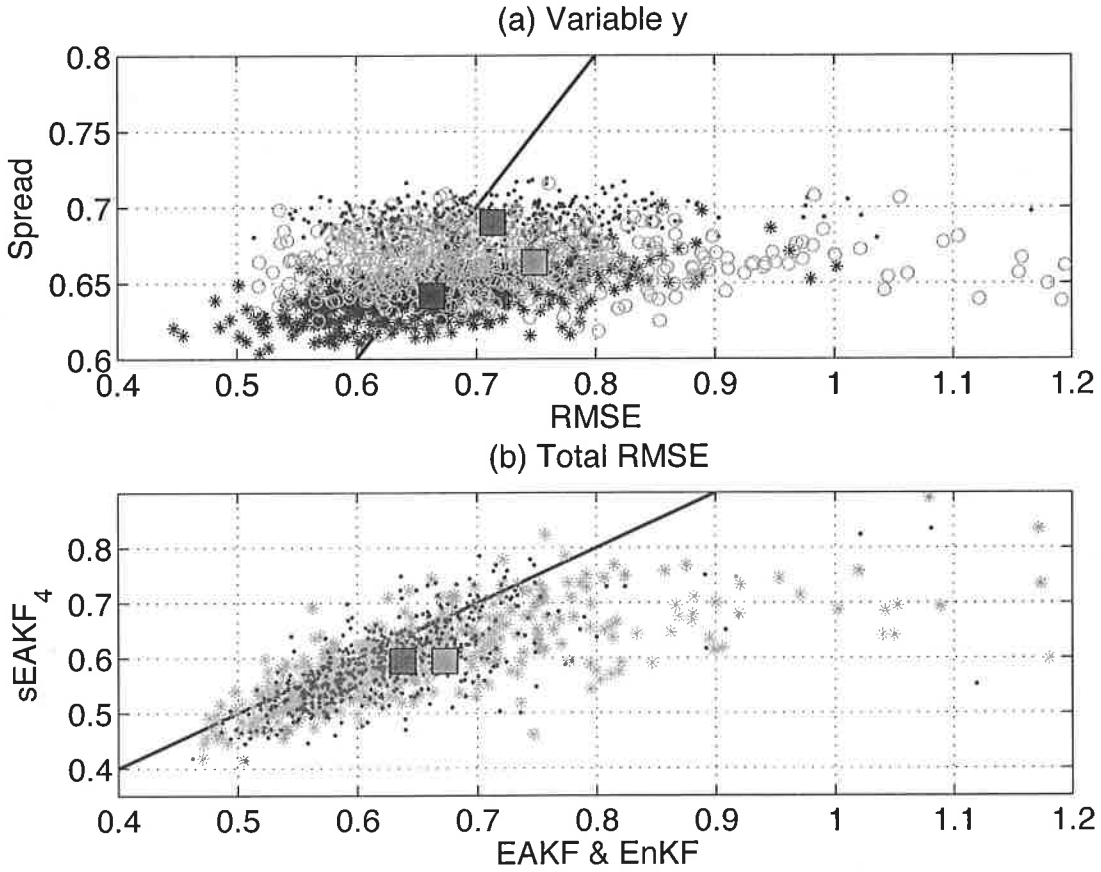

Figure 4 The same as figure3 except for $sEAKF_4$ experiments with a 20-member ensemble and the corresponding EnKF (EAKF) simulations.

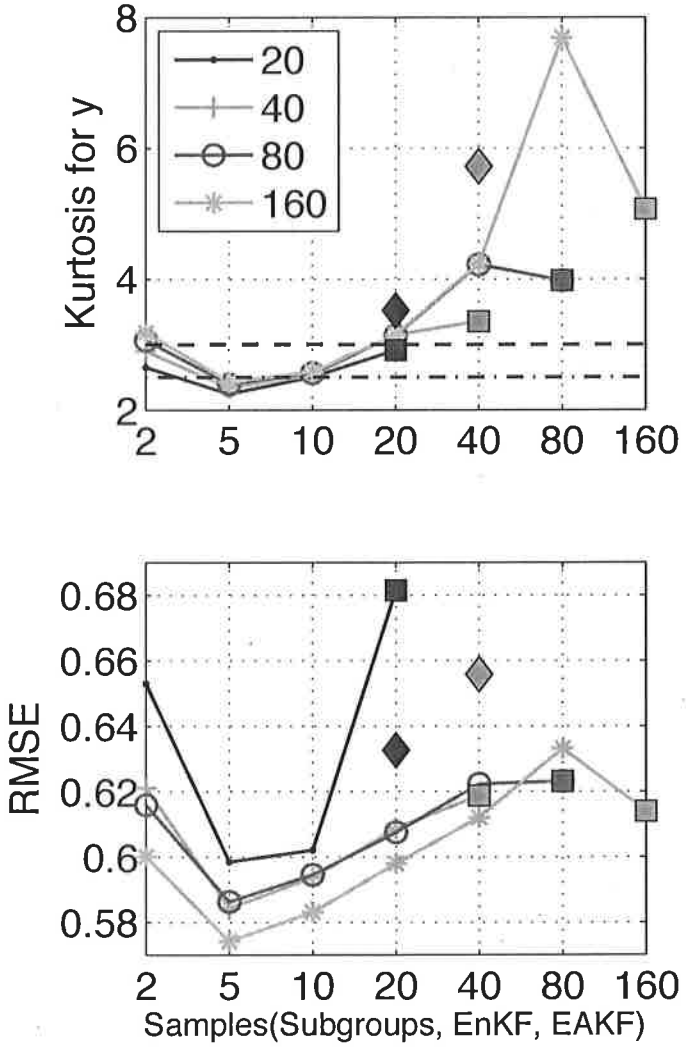

Figure 5 The ensemble kurtosis and analysis RMSE for different ensemble size averaged from 500 experiments.

The x-axis represents the sample sizes for subgroups of $sEAKF_n$ (EnKF, EAKF). The blue dot lines are for 20-member ensembles; green plus lines are for 40-member ensembles; red circle lines are for 80-member ensembles; and cyan star lines are for 160-member ensembles. The squares represent the results from EnKF simulations and the diamonds

represent the average results from EAKF simulations. The two black dish lines on upper panel are the kurtosis for Lorenz63 system (2.5) and for a Gaussian white noise distribution (3).

The kurtosis (RMSE) from EAKF for an ensemble size of 80 and 160 are 14.5 (0.75) and 56.8 (1.05) that are too large to be shown on the plots.

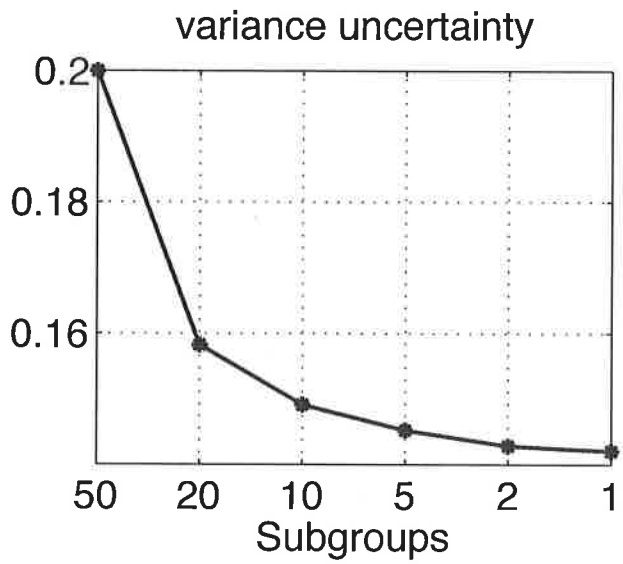

Figure 6 The variance uncertainty generated by limited sampling and different subgrouping for a Gaussian PDF with variance 1. The variance uncertainty derives from 100,000 Monte Carlo realizations.