

A Random Subgrouping Scheme for Ensemble Based Filter (sEnSRF)

Yun Liu^{*1}, Xingyao Rong^{1,2}, Zhengyu Liu^{1,3}, Shu Wu¹, Shaoqing Zhang⁴, Robert Jacob⁵

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 filter and deterministic filter. Both suffer an outlier problem when they are applied to a nonlinear system, especially for the deterministic filter. A nonlinear system generates outliers in an ensemble. For a deterministic filter, the outliers can persist for a long time and develop into extreme outliers, which tend to generate large analysis errors.

To address this problem, a random subgrouping technique is developed here to overcome the effect of outliers in deterministic filters. The new technique uses deterministic filter algebra but adds stochastic information into the filter system through random subgrouping. Test results using random subgrouping technique with two low-order models (Lorenz-63 and Lorenz-95) show that the new scheme dramatically improves performance compare 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 ensemble based filter is to derive the forecast uncertainty from an ensemble of model integrations. The ensemble based filter can be divided into two categories: stochastic filter and deterministic filter. 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 (Anderson 2001, Bishop et al. 2001, Whitaker and Hamill 2002, Tippett et al 2003, Sakov and Oke 2008), also called an Ensemble Square Root Filter (EnSRF), transforms

the ensemble anomaly to match the theoretical variance given by the Kalman Filter (KF) theory (Kalman 1960), while a stochastic filter (EnKF) attempts to match the updated variance with the KF theory by adding perturbations to the observations (Burger 1998, Houtekamer and Mitchell 1998, Evensen 2003).

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 shows in both background uncertainty and observational uncertainty. An EnSRF avoids the sampling error introduced by perturbed observation and tends to generate better analysis than an EnKF in a linear model especially for a small ensemble size (~10-20) (Whitaker and Hamill 2002, Evensen 2003, Anderson 2010).

However, an EnKF performs better than EnSRF when they face the challenges from non-Gaussian PDFs (Lawsen and Hansen 2004, Lei et al 2010, Anderson 2010). A nonlinear system tends to generate outliers in an ensemble. An EnKF can mix the outliers by adding random noise to the ensemble members through perturbed observation. Therefore the effect of outliers is relatively weak. An EnSRF tends to have persistent outliers because it has no effective way to “pull” the outliers back to the right track. A persistent outlier can therefore drift afar to become an extreme outlier, leading to a large analysis error or filter divergence (Lawsen and Hansen 2004, Anderson 2010). Furthermore, a larger ensemble size will have more outliers, so an EnSRF for a nonlinear

system performs worse as the ensemble size increases (Lawsen and Hansen 2004, Mitchell and Houtekamer 2009, Anderson 2010).

To improve the performance of the ensemble based filter in a nonlinear system, the key is to eliminate the effect of outliers. Several methods have been proposed recently. For example, Sakov and Oke (2008) use a random transformation to prevent outliers in the EnSRF, while Anderson (2010) uses a rank histogram filter to eliminate outliers. Here 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 improves the filter analysis in a nonlinear system significantly. 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 merges the model forecasts with the observation statistically to generate analysis with a reduced error. Assume $\mathbf{x} = \mathbf{x}(t)$ is an n-dimensional column vector for the model state at time t and y_t is the observation at time t , data assimilation solves the conditional probability distribution density function (PDF) $P(\mathbf{x}(t)|Y_t)$ of model states at time t , where $Y_t = [y_1, y_2, y_3, \dots, y_{t-1}, y_t]$. Based on Bayes' rule, the conditional probability distribution can be written as

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

Kalman Filter

The KF can achieve an optimal estimation of $x(t)$ for either a minimum variance sense or a maximum likelihood sense with the assumption of error from Gaussian PDFs, which can be represented by the mean and covariance,

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

The KF updates the analysis conditional mean (x^a) and variance (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 which is weighting coefficient matrix related to the uncertainties of forecast and observation. The H is the linearized mapping function from state variables to observation space, $y=Hx$. The observational uncertainty R is given by observation itself. The forecast uncertainty P^f is advanced from previous P^a by using Kolmogorov equation.

EnKF

Different from KF, ensemble method uses a Monte Carlo method to generate an ensemble x_i^a to sample $P(x(t)|Y_t)$ and integrates the ensemble to achieve x_i^f which resolves $P(\mathbf{x}(t+1)|Y_t)$, where $i=1,2,\dots,N$. The observations are treated as random variables that are perturbed to sample the uncertainty of observations, so the analysis variance will match what derived theoretically in KF in eqn. (4) (Burgers et al 1998, Houtekamer and Mitchell 1998).

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

EnSRF

EnSRF transforms the forecast ensemble to match the analysis error as in (3) and (4). This paper will show the results of Ensemble adjustment filter (EAKF, Anderson 2001, 2003). Similar results are also obtained for subgrouping schemes on other EnSRFs, such as the ensemble transform filter (ETKF, Bishop et al. 2001). As one type of EnSRF, EAKF updates the ensemble in two steps (Anderson 2003). First it derives the analysis ensemble mean and variance and computes the ensemble increment for the ensemble to match the analysis error in the observation space. Then, the ensemble increment is distributed over relevant state variables through a least square fitting.

sEnSRF

What is to be introduced is a simple, yet new filter scheme called random subgrouping EnSRF (sEnSRF). This scheme is the same as a regular EnSRF except that the entire ensemble is sub-grouped randomly at each analysis step. In sEnSRF, at each analysis step, the entire ensemble is divided into sub-ensembles of equal size randomly; all the sub-ensembles are updated independently using the same observation, but its own background covariance. Each sub-ensemble will have a different combination of ensemble members at different analysis time and the analysis and forecast are constructed using the entire ensemble. Therefore, sEnSRF shares the same algebra as EnSRF, but adds some stochastic variability through the random subgrouping. This random subgrouping helps to eliminate the persistent outliers because any ensemble member will be grouped into

different sub-ensembles at different analysis times randomly and therefore will have little chance to drift away persistently too far from the ensemble mean analysis. The objective of our sub-grouping therefore differs from that of Houtekamer and Mitchell (1998), where a sub-grouping is used to reduce the negative bias in the analysis error variance. Here, the random subgrouping is used to eliminate the effect of persistent outliers that often occur in a regular EnSRF. Therefore, sEnSRF is still a square root method but no longer deterministic. Specifically, sEnSRF can be performed generically in 4 steps, as shown schematically in Figure 1:

- (1) the model ensemble is integrated forward until the observation arrives;
- (2) the N -member ensemble is divided into n sub-ensembles, randomly, each of N/n members;
- (3) each sub-ensemble is updated independently using EnSRF with the same observation;
- (4) steps 1 to 3 are repeated until the all the observation arrives.

It is conceivable that sEnSRF can perform 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 observation in EnKF.

3. sEnSRF in the Lorenz model

A) The sEnSRF Scheme

We first apply sEnSRF to EAKF in the Lorenz63 model (Appendix A). Hereafter, a random subgrouping EAKF into n sub-groups will be denoted as sEnSRF _{n} . Our sEnSRF _{n} will be compared with the EnKF and EAKF on the same system using the same observation and same initial condition.

Due to the strong nonlinear nature of the Lorenz63 model, it is easy to produce outliers in a 100-member ensemble in EnKF or EnSRF simulation, but not in SEAKF₁₀ (Fig. 2). In EAKF (Fig.2b), an extreme outlier persists for a long time, with a deviation about one order larger than the ensemble spread. This occurs because of the lack of mechanism to pull it back towards the ensemble mean. An EAKF tends to retain high-order moments through the assimilation process (Anderson 2001), which may be a good choice for filter problem. However, this feature also leads to outliers persistent during the EAKF assimilation process. The persistent outliers can be separated far away from most other members to become extreme outliers, as shown in Fig. 2b.

The outliers are less extreme in EnKF (Fig.2a) than in EAKF (Fig.2b). This is because the random perturbation on observation plays the role to adding random noise to outliers in EnKF. Therefore, an outlier can't drift too far away from the ensemble mean before being pulled back after a few cycles of analyses.

Corresponding to the EAKF, the random subgrouping of 10 sub-groups (sEAKF₁₀) is performed. The random subgrouping plays a similar role as the perturbed observation does in EnKF, so no outliers persist during the assimilation (Fig.2c). An ensemble member that is an outlier in one subgroup combination may be a regular member for another subgroup combination in the next step. As a result, an outlier at one analysis step can be eliminated by the random subgrouping at the next analysis step. Furthermore, the sEnAF₁₀ has less chance to produce outlier than EnKF and EAKF because each sub-ensemble has much smaller sample size than the full ensemble and a small sample size produces less outliers.

The reduced outlier can be quantified through the ensemble kurtosis which is a

good index for the presence of outliers

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

where \bar{x} is ensemble mean. Theoretically the kurtosis of a Gaussian distribution is 3. The kurtosis for the y -variable in the Lorenz63 system is calculated from a long control run as being 2.5. The time-mean kurtosis of y is ~ 20 for the EAKF with a 100-member ensemble. This is much larger than the kurtosis of the Gaussian distribution. The time-mean kurtosis of y is ~ 4 for EnKF with a 100-member ensemble, indicating a weak effect of outliers relative to Gaussian. In comparison, the kurtosis of y in sEAKF₁₀ is 2.6. This is very close to the kurtosis of the Lorenz63 system and indicates little spurious outliers generated by the assimilation scheme. Therefore, sEAKF₁₀ suffers little from the outlier problem relative to EnKF and, especially, EAKF. In other words, the sEAKF₁₀ retains statistically the 4th-order moment while the EAKF can't.

The sEAKF₁₀ also generates the smallest analysis error among the three filter schemes, because of its lowest kurtosis. sEAKF₁₀ reduces the root mean square error (RMSE) of the analysis to 0.66 for the variable y , which is much smaller than the RMSE for EAKF (0.95) and better than the EnKF (0.69). Statistically, sEAKF₁₀ and EAKF simulations show the consistence between RMSE and ensemble spread (figure 3b), though the ensemble spreads is a slightly smaller than the RMSE (figure 3b).

Because the persistent outliers distort PDF, some EAKF simulations generate very big analysis error (with the RMSE > 2) but very small ensemble spread (~ 0.7). In a few cases, the EAKF even becomes filter divergent (Fig. 3a). Starting from the same first guess (initial conditions) and using the same observations, more than 95% (80%) sEAKF₁₀

experiments produce smaller RMSE than EAKF (EnKF).

B) Optimal size of the sub-group

The number of subgroups (n) is a free parameter in $sEAKF_n$. The question then arises: what is the best subgroup number? The random subgrouping scheme eliminates the effect of outliers by introducing randomness into the filter system. To fulfill this goal, $sEAKF_n$ requires enough stochastic freedoms and small enough sub-ensemble size. If the subgroup is too small, the effect of outlier still exists though it is reduced significantly. For example, the ensemble kurtosis of $sEAKF_2$ is ~ 4.8 for the 100-member ensemble experiments and ~ 9.8 for 200-member ensemble experiments (Fig. 4a). However, the numbers of subgroup is bounded by the ensemble size. The biggest subgroup is $N/2$ for a N -member ensemble.

In the mean time, 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 all converges to the kurtosis of the Gaussian distribution 3, instead of the truth kurtosis 2.5 (Fig.4a). This occurs because each subgroup has only two members, making it impossible to resolve a PDF and the high-order moments beyond that of Gaussian. (The Gaussian results from the average of a large number of subgroups). Therefore, both 2th-order and the 4th-order moments are distorted.

A larger ensemble size has more probability to produce outliers and therefore requires more subgroups to eliminate outliers. In the mean time, a larger ensemble size can accommodate more subgroups. As a result, for different sizes of the full ensemble, the minimum kurtosis seems to be achieved at about the same size of sub-ensemble

(Fig.4a), which corresponds to different numbers of subgroups. For the full ensemble sizes of 50, 100 and 200, the minimum kurtosis all converges to the system kurtosis 2.5 at the sub-ensemble size 5 (Fig. 4a). This is the best subgrouping with the forth-order moment conserved and the effect of outliers removed.

It is conceivable that a reduced error in sampling kurtosis should reduce the outliers and therefore improve the filter performance. This is indeed the case as seen in the corresponding RMSE (Fig.4b). Overall, the filter performance largely follows the sample kurtosis for our system. Smaller RMSE occurs for smaller sample kurtosis. For the 100-member ensemble size, sEAKF₂₀ gives the smallest analysis RMSE, which is reduced by 30% from that of the standard EAKF. This size of sub-sample 5, with the best sample kurtosis, also appears to be the optimal sub-sample size with the smallest RMSE for other sizes of the full ensemble.

C) Sampling error for subgrouping

Our random subgrouping scheme successfully removes the outliers and maintains the high-order moments, but it also introduces sampling errors into the filter simulation. Since each sub-ensemble has a much smaller sample size than the full ensemble, the background uncertainty from an individual sub-ensemble has a greater sampling error than that of the full ensemble. However, the PDFs constructed from all the sub-ensembles automatically represent the uncertainty of the forecast error PDFs, which leads to a compensation of the sampling error generated by the subgrouping. Therefore the net increase in sampling error can still be limited.

A Mont Carlo method is used to evaluate the sampling error caused by subgrouping (Fig. 5). 100 samples are used to resolve a Gaussian PDF. The subgrouping

scheme is also implied during variance calculation. The total sample variance is the average variance of each subgroup. For example, the 100 samples are divided into n groups of $100/n$ samples and calculate the variance independently for each subgroup and then average them to get the total variance. The variance uncertainty produced by limited sampling is represented by the standard deviation of variances from 100,000 Monte Carlo experiments.

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

D) Subgrouping in EnKF

An EnKF can also benefit 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 100-member ensemble experiments and ~ 6 for 200-member ensemble experiments (figure 4). The subgrouping scheme is applied in the EnKF 100-member ensemble experiments. The analysis RMSE decreases 4% comparing with EnKF and the kurtosis decreases to ~ 3 when the 5 sub-ensemble are used (figure not show).

E) sEnSRF in Lorenz 96 model

The improvement of analysis by sEnSRF can also be shown in the Lorenz96 model (Lorenz, 1996) (Appendix B). In addition, in Lorenz96 model, we can also study the effectiveness of sEnSRF to systems of different levels of chaos. (Table 1). When the

system is near neutral ($F=2$), the RMSE are comparable in sEAKFn and EAKF, implying an insensitive to the subgrouping scheme. As the system becomes more chaotic, sEAKFn performs better than EAKF. For the strongly chaotic cases ($F=8, 10$), the RMSE is reduced by $\sim 10\%$ in sEAKF₁₀ than in EAKF. Similar to the low dimension case in Lorenz63 model, sEAKFn performs better as the number of subgroups increases.

4. Summary

The persistent outlier problem arising from non-Gaussian PDFs is a challenge for ensemble based filter, especially for EnSRF of large ensemble size (Lawson and Hansen 2004, Anderson 2010, Lei et al 2010). The sEnSRF scheme solves this problem with a random subgrouping in EnSRF, improving the data assimilation quality in nonlinear systems significantly. The sEnSRF uses the formula of EnSRF for each subgroup, but is no longer a deterministic filter. It avoids sampling errors introduced by the perturbation on observations and eliminates the effect of outliers. It also can retain the high-order moments through the assimilation process.

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

sEnSRF is applied to two simple models: the Lorenz63 model and the Lorenz96 model. The random subgrouping improves the filter analysis significantly relative to both the stochastic filter EnKF and the deterministic filter EnSRF. Comparing with deterministic filter, it can decrease 10 \sim 30% of analysis error under strong chaotic

conditions.

One advantage of sEnSRF is its simplicity and practicality. sEnSRF can be easily applied to a high dimension system. It is particular effective in highly chaotic systems which tend to generate more extreme outliers. Therefore, we propose sEnSRF as an effective and practical assimilation method for complex weather and climate models.

Table 1 The mean analysis RMSE for sEAKF_n and EAKF on Lorenz95 system from 200 realizations. The first column denotes the forcing term in the system.

	sEAKF ₁₀	sEAKF ₅	sEAKF ₂	EAKF
F=2	0.0253	0.0247	0.0251	0.0251
F=5	0.1424	0.1445	0.1545	0.1545
F=8	0.3131	0.3147	0.3487	0.3492
F=10	0.3745	0.3774	0.4217	0.4227

Appendix A Lorenz model

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

$$\begin{aligned}\dot{x} &= -\sigma(y - x) \\ \dot{y} &= \beta x - y - xz \\ \dot{z} &= xy - cz\end{aligned}$$

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 ($\beta = 28$, $\sigma = 10$, $c = 8/3$). The evolution of the state vector (x, y, z) describes 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 1000 time units and the “observation” by adding on the “truth” random errors with standard deviation $(2, 2, 2)$.

For most of the experiments, the observation time interval use 0.1 and ensemble size use 100. The inflation scheme was not used in all the experiments for the fare comparison.

Appendix B

Lorenz 96 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 $i = 1, \dots, N$ with a cyclic indices. The N equals to 200 for our simulations. The model is integrated using a 4-th order Runge-Kutta method with a time resolution of $dt=0.01$. To investigate the filter performance under different conditions, the forcing term F choose 2, 5, 8, 10, which let the system shift from almost linear to strong chaos.

All the simulations use ensemble size 100 and a perfect observation system with observation frequency 0.1 (10 time steps) and the error standard derivation 1. The influence radius for localization is 11 for all the simulation. The inflation scheme was used in all the experiments

Figure 1 Flow chart of the SEnSRF data assimilation system for the case of 25-member ensemble and 5 subgroups. Each small square denotes an ensemble member and each line denotes a sub-ensemble.

Figure 2 Initial error evolution of 100 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 SEAKF10 simulation. (The figure follows figure 3 in Anderson 2010).

Figure 3 The scatter diagram of analysis RMSE and ensemble spread of variable y for different filter schemes. Each filter scheme has 200 realizations. The red dots are for EAKF simulation, green circles are for EnKF and blue stars are for sEAKF₁₀.

The lower panel is the fine scale of upper panel. The squares represent the average of total 200 experiments, red for EAKF, green for EnKF and blue for sEAKF₁₀. The black line denotes that the RMSE equals to ensemble spread.

Figure 4 The kurtosis and analysis RMSE for different ensemble size are averaged from 200 realizations.

The x coordinator represents the samples for each sub-ensemble. The blue dot lines are for ensemble size 50; green plus lines are for ensemble size 100; and red circle lines are for ensemble size 200. The squares represent the results from EnKF simulations. Two black dash lines on upper panel are the kurtosis for Lorenz system (2.5) and for Gaussian white distribution (3).

The kurtosis and RMSE from EAKF are no shown on the plot because they are much bigger than the results from EnKF and sEAKFn.

Figure 5 The variance uncertainty generated by limited sampling and different subgroups for a Gaussian PDF with variance 1. The variance uncertainty derives from 100,000 Mont Carlo realizations.

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.
- 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 filters as 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: Predictability: a problem partly solved. In: *Proceedings of the ECMWF seminar on predictability*, vol1, ECMWF, Reading, United Kingdom, 1-18.
- Mitchell, H. L. and P. L. Houtekamer, 2009: Ensemble Kalman Filter Configurations and Their Performance with the Logistic Map. *Monthly Weather Review*, 137, 4325-4343.
- 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 perturbed observations. *Monthly Weather Review*, 130, 1913–1924.

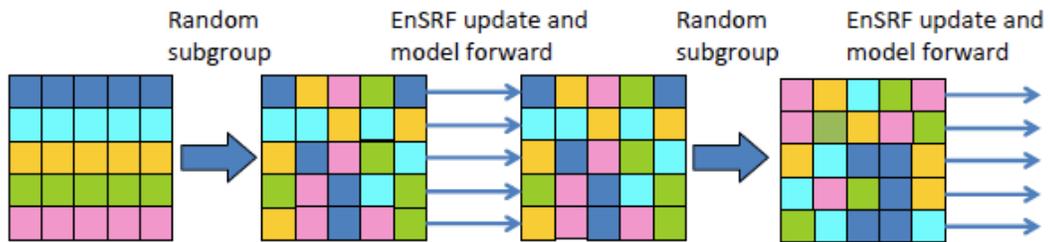


Figure 1 Flow chart of the SEnSRF data assimilation system for the case of 25-member ensemble and 5 subgroups. Each small square demotes an ensemble member and each line denotes a sub-ensemble.

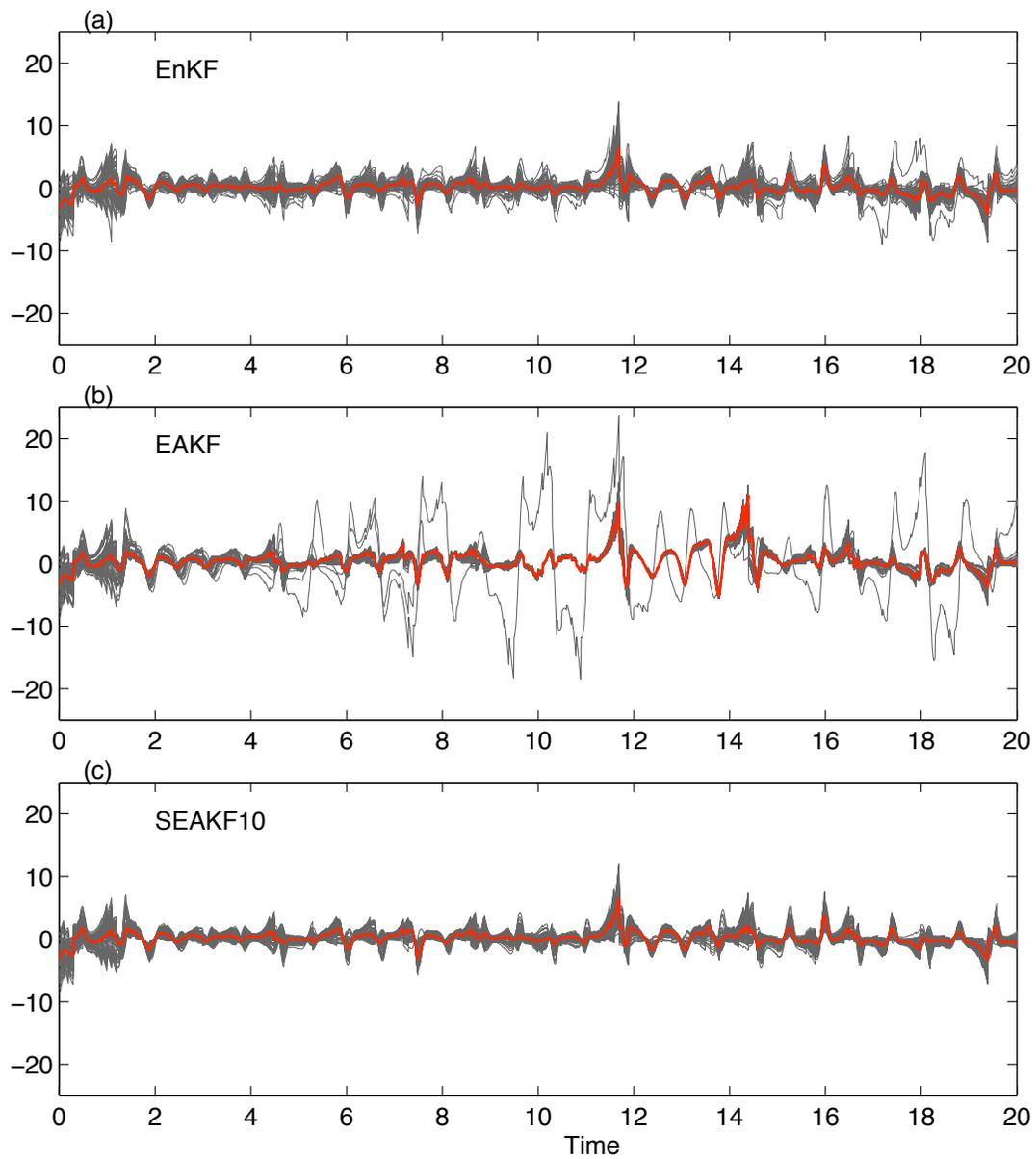


Figure 2 Initial error evolution of 100 ensemble members (black lines) and the ensemble means (red lines) for variable y in Lorenz system. (a) is for EnKF simulation; (b) is for EAKF simulation and (c) is for SEAKF₁₀ simulation. (The figure follows figure 3 in Anderson 2010).

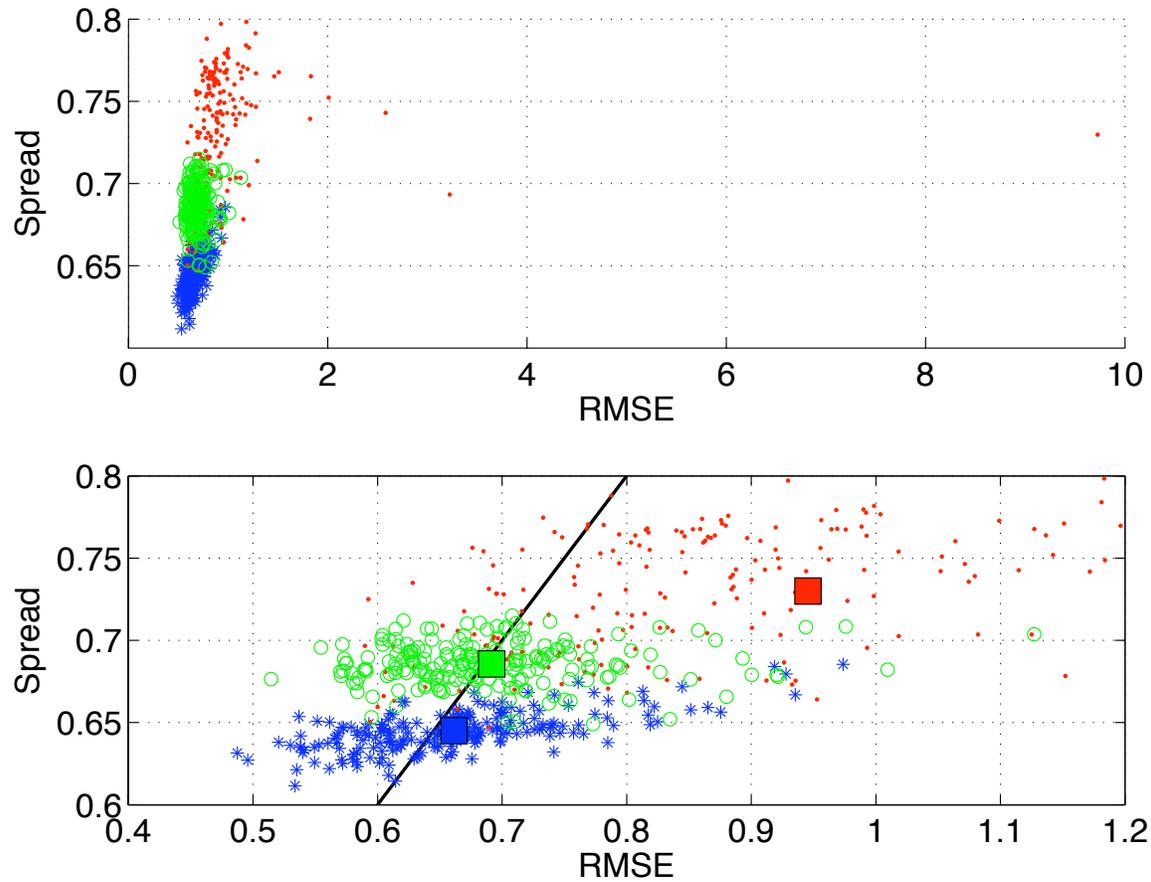


Figure 3 The scatter diagram of analysis RMSE and ensemble spread of variable y for different filter schemes. Each filter scheme has 200 realizations. The red dots are for EAKF simulation, green circles are for EnKF and blue stars are for sEAKF₁₀. The lower panel is the fine scale of upper panel. The squares represent the average of total 200 experiments, red for EAKF, green for EnKF and blue for sEAKF₁₀. The black line denotes that the RMSE equals to ensemble spread.

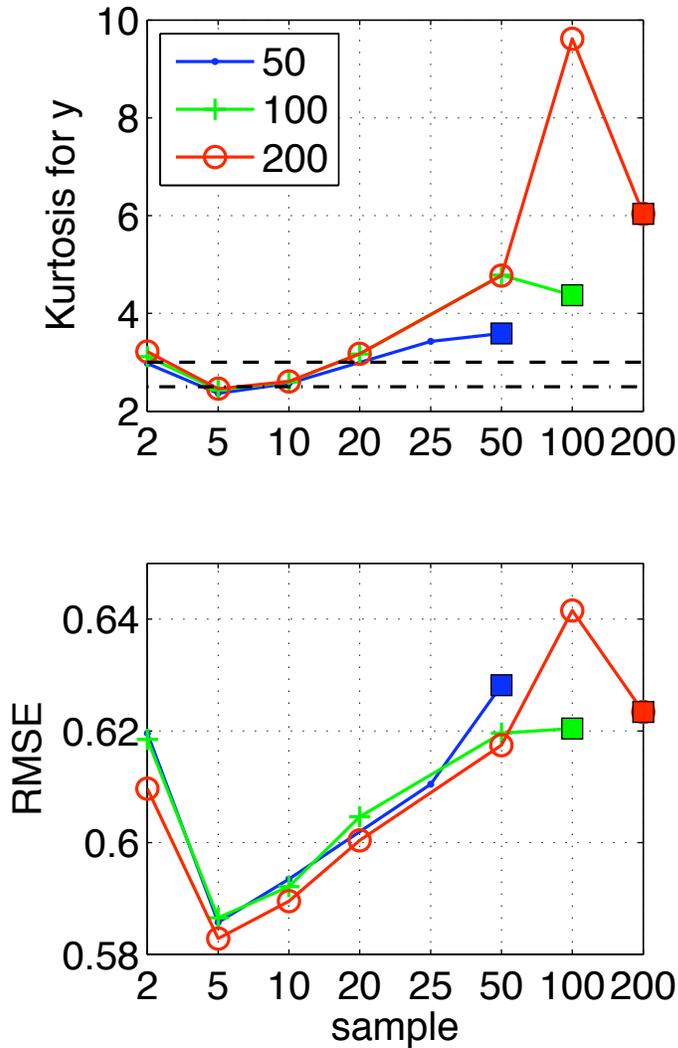


Figure 4 The kurtosis and analysis RMSE for different ensemble size are averaged from 200 realizations.

The x coordinator represents the samples for each sub-ensemble. The blue dot lines are for ensemble size 50; green plus lines are for ensemble size 100; and red circle lines are for ensemble size 200. The squares represent the results from EnKF simulations. Two black dash lines on upper panel are the kurtosis for Lorenz system (2.5) and for Gaussian white distribution (3).

The kurtosis and RMSE from EAKF are no shown on the plot because they are much bigger than the results from EnKF and sEAKF_n.

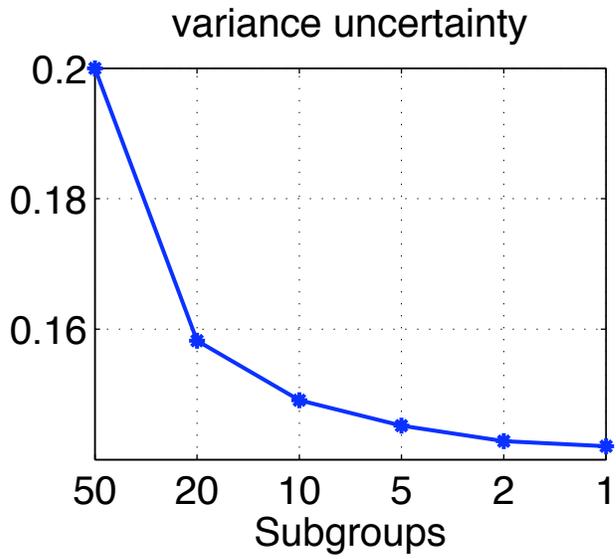


Figure 5 The variance uncertainty generated by limited sampling and different subgroups for a Gaussian PDF with variance 1. The variance uncertainty derives from 100,000 Mont Carlo realizations.