P 6.13

A COMPARISON OF DIFFERENT ENSEMBLE GENERATION TECHNIQUES

Brian J. Etherton and Craig H. Bishop
The Pennsylvania State University

1. INTRODUCTION

Observations from the standard global observing network, such as weather balloons, satellite observations, and surface reports, are incorporated into forecast models through data assimilation. Data assimilation combines information in observations with information in a first guess field to produce a new analysis. The first guess field is usually the forecast from a numerical weather prediction model valid at the time the observations are taken. Data assimilation blends the observations with the first guess field using estimation theory, such as maximal likelihood estimation and minimum variance estimation (Daley 1997). To use estimation theory for data assimilation, error statistics for the observation error and the first guess error are required. These error statistics are used to minimize the error variance of the new analysis and thus provide the best possible approximation of the state of the atmosphere.

Error statistics for observations are based on the precision of the instrument, and are provided by the manufacturer. There are also error statistics for the error of representation, an error associated with how one piece of data may not accurately represent the air mass it is located in. An example is a weather balloon that ascends through a thunderstorm while the surrounding air is storm free. The observed temperature and moisture vertical profile is not indicative of the adjacent environment, as the scale of the feature observed is not the scale of the region being represented by the weather balloon. Finally, there are the error statistics associated with the first guess field.

The error statistics for the first guess field currently used in most operational centers parameterized from a long time series of previous forecasts, and are generally spatially homogeneous and temporally invariant (Parrish and Derber, 1992). These fixed statistics cannot account for the differences between atmospheric conditions. For example, the error statistics near a cold front are likely to be very different to those in the center of a high-pressure system. To produce the analysis with the analysis with the smallest error variance, the optimal error statistics are needed. Given the shortcomings of the error statistics currently used in many operational centers, better-suited error statistics are proposed. The focus of this research is to use ensembles to generate error statistics, and then try to determine which ensemble generation method produces the best error statistics.

An approach to generate these error statistics is a hybrid ensemble transform Kalman filter / 3D-Var data assimilation scheme. This scheme was introduced in Hamill and Snyder (2001) and applied in Bishop et al (2001). The hybrid approach makes a move towards flow dependent error statistics while not abandoning the robust conventional approaches used most often in operations.

2. THE HYBRID ENSEMBLE TRANSFORM KALMAN FILTER / 3D-VARIATIONAL DATA ASSIMILATION SCHEME

The Ensemble Transform Kalman Filter (ET KF) (Bishop et al., 2001) is a method of blending observations and a first guess field into a model analysis using an ensemble-based estimate of the error statistics. The Kalman Filter (Kalman and Bucy, 1961) is the optimal means of spreading out information to a model. The basic equation for the Kalman gain matrix, $\mathbf{K}(t_f,t_i)$ is:

$$\mathbf{K}(t_f,t_i) = \mathbf{P}^f(t_f,t_i)\mathbf{H}^T\,(\mathbf{H}\mathbf{P}^f(t_i,t_i)\mathbf{H}^T+\mathbf{R})^{-1}$$

Where $\mathbf{P}^f(t_f,t_i)$ is the prediction error covariance matrix correlating errors at time t_f with errors at time t_i, **R** the observation error covariance matrix, and **H** a matrix that interpolates from observation space to model space given observations. The time t_i is the time that the observations are taken, the time t_f when the time the prediction is valid at. The Kalman filter is the optimal method of combining a first guess field with observations given that $\mathbf{P}(t_f,t_i)$ and **R** are known exactly. **R** is a diagonal matrix with values along the diagonal equal to the variance of the observation error, which is prescribed. The focus of the hybrid scheme is the generation of the error statistics associated with the first guess field, $\mathbf{P}^f(t_i,t_i)$.

Using a standard 3D-var approach, the prediction error covariance matrix is approximated using a parameterization. This covariance matrix is denoted as **B**. This matrix is invariant, homogeneous, and generally based on some sort of climatology. For this study, an approach similar to that in Parrish and Derber (1991) was used. To improve upon these error statistics, a flow

Corresponding author address: Brian J. Etherton, Penn State University, Department of Meteorology, University Park PA, 16802; e-mail etherton@essc.psu.edu

dependent covariance matrix is produced by taking the outer product of ensemble members. This matrix is named $\mathbf{F}^f(t_i,t_i)$, and $\mathbf{F}^f(t_i,t_i) = \mathbf{X}_i\mathbf{X}_i^T$, where the "f" subscript denotes that the ensemble perturbations and the error statistics are valid at the forecast time.

For the hybrid ensemble Kalman filter / 3D-Var scheme, the prediction error covariance matrix is set equal to a combination of both climatological and flow dependent covariances:

$$\mathbf{P}^f(t_i,t_i) = \alpha\mathbf{F}^f(t_i,t_i) + (1-\alpha)\mathbf{B} \qquad (2)$$

The main challenge with the Kalman filter is the inversion of the matrix $\mathbf{HP}^f(t_i,t_i)\mathbf{H}^T+\mathbf{R}$. Using the ensemble based covariances, this inversion is not as computationally expensive, as the rank of $\mathbf{HP}^f(t_i,t_i)\mathbf{H}^T+\mathbf{R}$ is no larger than the number of ensemble perturbations, whereas the full $\mathbf{HP}^f(t_i,t_i)\mathbf{H}^T+\mathbf{R}$ has a rank equal to the number of observations. For a global analysis scheme, this can be quite large. For the 3D-Var part of the hybrid scheme, the inverse is calculated once, and stored on disk. For the inverse of the ensemble based covariance matrix, calculations are performed for every analysis cycle. Implicit in this inversion is that the ensemble perturbations are orthogonal to each other, such that the matrix is not rank deficient.

An application of the hybrid ensemble transform Kalman filter / 3D-var scheme was done on a quasi-geostrophic model in Hamill and Snyder (2000). In their study, the method of generating the ensemble was to have separate analysis cycles for each ensemble member. This approach, referred to as "perturbed observations" did produce decent ensemble perturbations, but was costly, in that a separate analysis had to be produced for each ensemble member.

3. TYPES OF MODEL ERRORS

For the experiment, the same barotropic vorticity model used to generate the truth, and hence, the "observations", was used as the forecast model. This simple model has only the effect of beta and a relaxation scheme as forcings. To show the ability of the data assimilation schemes to deal with different situations, there were three different approaches to the truth run. The first option was to have the model and the truth run be the same model, same resolution, and same vorticity field the relaxation scheme nudges the model towards. In this case, it is expected that after repeated sampling of the model domain, that the errors in the first guess field will diminish to near zero.

The second option was to have the vorticity relaxation scheme of the model relax to a different field than was relaxed to in the truth run. This was done by having the model relax to its initial vorticity value along its leftmost gridpoint for each latitude. The truth run relaxed to the initial conditions. No term representing the model error was included in the Kalman Filter equations, as the only two covariance matrices were the prediction error covariance matrix, and an observation error covariance matrix. The observation error covariance matrix consisted of a diagonal matrix with the values along the diagonal equal to the variance associated with the random values (mean 0, normal distribution) added to values from the truth run when an observation was taken.

The third option was to run the truth at a higher resolution than the model was run at. This was accomplished by running the truth at 100km, while the model itself was run at 200km. The third was to run the truth at the same resolution as the model, but to have the vorticity relaxation different between the two runs. In this way, the forcing of the model was different to the forcing of the truth, creating a model error.

4. ENSEMBLE GENERATION SCHEMES IN THIS STUDY

Perturbed Observations

Developed in Houtekamer and Derome (1995), this approach uses 17 independent model runs, rather than one control run and 16 perturbations. Each model run is treated as a first guess, and has it's own analysis increments made. Error covariances are computed using the 16 ensemble members other than the first guess. To represent observation error uncertainty in the ensemble, the innovation vector ($\mathbf{d}-\mathbf{Hx}(t_i)$) has additional random errors added to it. These errors are in accordance with the values in the error covariance matrix, **R**. After the full collection of ensemble members has a new analysis, the control run is set to the mean of the 17 analysis fields.

With only difference in the observation error between the ensemble members, the members would drift to resembling each other at the analysis cycles continued. To maintain ensemble spread, a rescaling of the ensemble perturbations is done using a maximal likelihood estimation theory developed in Dee (1995). In this approach, it is expected that the square of the projection of the innovation vector ($\mathbf{d}-\mathbf{Hx}(t_i)$) onto the eigenvectors of $(\mathbf{HP}(t_i,t_i)\mathbf{H}^T+\mathbf{R})^{-1}$ will be equal to the spread of the matrix $(\mathbf{HP}(t_i,t_i)\mathbf{H}^T+\mathbf{R})^{-1}$ A rescaling parameter is calculated such that the sum of the eigenvectors of $\mathbf{HP}(t_i,t_i)\mathbf{H}$, when multiplied by it, are equal to the sum of the squared projected innovations minus the eigenvectors of **R**. These rescaled perturbations are then added back onto the ensemble mean to generate the 17 different first guess fields. From there, the analyses for each member are produced.

Breeding Method

An attempt to reproduce the approach of Toth and Kalnay (1993), the 24 hour old ensemble of perturbations generated from the "random sample **B**" approach (explained later) are rescaled such that the sum of the squares of the singular values of the perturbations is equal to the sum of the squares of the climatological **B** matrix. These rescaled perturbations are then used as the initial conditions for an ensemble run.

Optimal Perturbations

An attempt to simulate the singular vector perturbations of Buizza and Palmer (1995). A 256-member ensemble is generated, using the top 256 directions of the climatological **B** matrix. Only the top 256 directions were used because these directions contained about 99 percent of the total variance in B, at one-quarter the computational cost to integrate all the eigenvectors of the B matrix. In essence, this approach attempts to propagate the entire phase space of the model forward in time. By integrating all (important) directions forward in time, the propagator matrix **L**, is constructed explicitly such that $\mathbf{x}(t_0) = \mathbf{L}\mathbf{x}(t_i)$, where $\mathbf{x}(t_0)$ is the initial conditions from a previous analysis, $\mathbf{x}(t_i)$ is the model first guess field at the time an increment is to be made. At the time the ensemble is used to make an increment, a singular vector decomposition is performed to find the top 16 directions. These are then used as the 16 perturbations for producing the error statistics of the first guess field.

Recycled $\mathbf{F}^a$

In this technique, the ET KF is used to calculate the analysis error covariance matrix, $\mathbf{F}^a(t_i,t_i)$, for each day. This matrix is then broken down into its eigenvectors and eigenvalues. Since $\mathbf{F}^a(t_i,t_i)$ was formed from a reduction of $\mathbf{F}^f(t_i,t_i)$, it is of the same rank as the number of ensemble members, hence, there are only 16 eigenvectors of this $\mathbf{F}^a(t_i,t_i)$ matrix (as opposed to the climatological $\mathbf{F}^a(t_i,t_i)$, which has the full 1024). These eigenvectors, multiplied by the square root of their eigenvalues, is used as perturbations to form the initial conditions for the ensemble members.

Top Directions of Parrish/Derber **B**

Using the Parrish/Derber derived covariance matrix, a singular value decomposition of **B** is performed, breaking the B down into 1024 singular vectors and 1024 singular values. From this, the top 16 directions are used, with each direction being multiplied by the square root of its singular value to form the 16 ensemble perturbations. Thus, the same set of perturbations is used every day.

With this ensemble generation technique, as well as all other schemes discussed in this section, a rescaling of the ensemble perturbations is done. The approach is similar to the approach for the perturbed observations, except the rescaling factor is applied to the next ensemble generation, rather than to the perturbations themselves.

Top Directions of Parrish/Derber $\mathbf{P}^a$

Using the equations developed for the ET KF, the climateological analysis error covariance matrix, $\mathbf{P}^a(t_i,t_i)$, is calculated using the equation $\mathbf{P}^a(t_i,t_i) = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T+\mathbf{R})^{-1}\mathbf{H}\mathbf{B}$. As with the climateological **B** Matrix, a singular vector decomposition of $\mathbf{P}^a(t_i,t_i)$ is performed, and again the top 16 directions of the $\mathbf{P}^a(t_i,t_i)$ matrix are multiplied by the square root of their singular values to form the ensemble perturbations. The **H** matrix used in this calculation is the same 144-observation site as was used in the generation of the **B** matrix.

Random Sample Parrish/Derber **B**

This technique uses a random sample of the climatological prediction error covariance matrix obtained using a method similar to that of Parrish and Derber (1991). A vector of random numbers, normally distributed (mean 0, standard deviation 1) are multiplied (element by element) to the singular values of the top 256 directions of **B**. The reason for using only the top 256 directions of **B**, as opposed to all 1024, is that the top 256 directions account for 99.8 per cent of the variance in **B**, so to include more simply wastes computer time. In the same way as when the top 16 eigenvectors of the **B** matrix were used, the daily rescaling of the initial perturbations is done to maintain an ensemble spread of a magnitude approximately the expected error of the first guess field.

Random Sample Parrish/Derber $\mathbf{P}^a$

Similar to the random sample of singular vectors of **B**, except the singular values and singular vectors of the climatological $\mathbf{P}^a(t_i,t_i)$ matrix are used. These structures are smaller scale than those produced from the **B** matrix, and are concentrated on the right (low observation density) side of the domain.

Gridpoint Perturbations

A Monte Carlo approach where random grid point vorticity perturbations are added to the first guess field. The random numbers come from a uniform distribution with mean zero, and maximum amplitude such that the sum of the squares of the perturbations was equal to the sum of the squares of the rescaled singular values of the top 16 directions of the climatological **B** matrix.

5. RESULTS

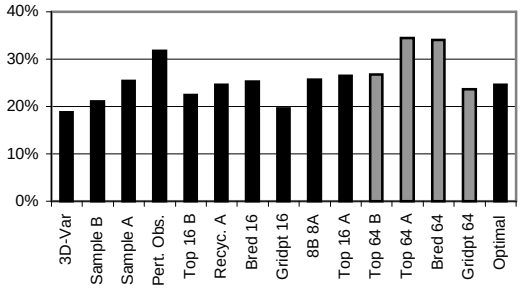
Increments were made using the hybrid ensemble Kalman filter / 3D-var data assimilation scheme, with an alpha value of 0.6, and with a 72 observation standard network. Results are the average for the first 25 days of the 99 day run, when the flow is the “least climatological”, and the ensembles provide the greatest improvement.

These results show the optimal perturbations method performs the best when the model error is the relaxation error, and the perturbed observations method performs the best when the resolution model error is present. However, several schemes perform with a similar skill, and are far less expensive to produce. For the resolution error case, with the exception of the grid point noise perturbations, that all schemes using a 16 member ensemble provide at least a 20 percent reduction of forecast errors, whereas pure 3D-var provides less than a 20 percent improvement on average.

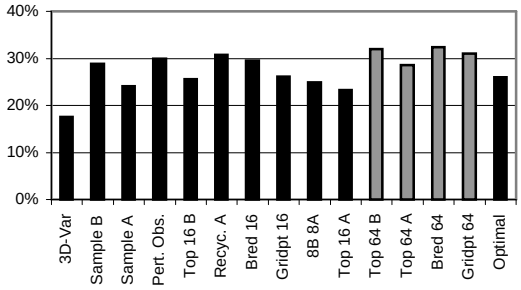
Given that the perturbed observations approach and the optimal perturbations approach take a great deal of computational effort to produce, it was hoped that they would provide far better analyses and forecasts. However, they are not much better than the very simple leading eigenvectors of the climateological **B** matrix. Perhaps a better use of computer power would be to form more ensemble members via this relatively cheap process. The idea being to attack the problem with bulk rather than precision.

The average daily percent reduction in global enstrophy 24 hour forecast error using a hybrid 3D-Var/ET KF data assimilation scheme

Resolution Error



Relaxation Error



Scores for 64 member ensembles constructed using the random sample of the climateological **B** matrix, the random sample of the climateological $\mathbf{P}^a$ matrix, the grid point perturbations, and the breeding method show that the breeding method and the top directions of the parameterized **B**-matrix have average daily improvements of about 30 percent for the relaxation error run. For the resolution error case, the top directions of climatological $\mathbf{P}^a$ matrix and the breeding method have better than 30 percent reduction in error. For each model error regime, these 64 member ensembles have greater error reductions than any of the methods used with only 16 ensemble members. Given the time that it took to build the optimal perturbations, or the separate analyses required in the perturbed observations approach, for this experiment, that time was better spent making more ensemble members via a simpler technique.

6. REFERENCES

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman Filter Part I: Theoretical aspects. Mon. Wea. Rev., 129, 420-435.

Borges, M. D., and D. L. Hartmann, 1992: Barotropic instability and optimal perturbations of observed zonal flows. J. Atmos. Sci.,49, 335-352.

Dee, D., 1995: On-line estimation of error covariance parameters for atmospheric data assimilaton. Mon. Wea. Rev., 123, 1128-1145.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter / 3d-variational analysis scheme. Mon. Wea. Rev., 128, 2905-2919.

Houtekamer P. L., and J. Derome, 1995: Methods for ensemble prediction. Mon. Wea. Rev., 123, 2181-2196.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. Mon. Wea. Rev., 126, 796-811.

Kalman R., and R. Bucy, 1961: New results in linear filtering and prediction theory. Trans. ASME, J. Basic Eng., 82D, 35-45.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorlogical Center's Spectral Statistical Interpolation Analysis System. Mon. Wea. Rev., 120, 1747-1763.

Toth, Z., and E. Kalnay, 1993: Ensemble Forecasting at NMC: The Generation of Perturbations. Bull. Amer. Met. Soc., 74, 2317-2330.