AN ENSEMBLE KALMAN SMOOTHER FOR REANALYSIS

Jeffrey S. Whitaker* and Gilbert P. Compo
NOAA-CIRES Climate Diagnostics Center, Boulder, CO

1. INTRODUCTION

Data assimilation is typically used to generate initial conditions for numerical weather forecasts. Therefore, each analysis is based upon only current and past observations. However, when producing a retrospective 'reanalysis', one is free to use all available observations, including those data collected after the analysis time.

A Kalman smoother is a direct generalization of the Kalman filter which incorporates observations both before and after the analysis time. Here we introduce the ensemble square-root smoother (EnSRS), which applies recent advances in the field of ensemble filtering to the fixed-lag Kalman smoother proposed by Cohn et al. (1994). The EnSRS uses Monte-Carlo estimates of forecast-analysis error cross-covariances needed to compute the Kalman smoother gain matrix. It is applied iteratively to a time series of observations, the first iteration is equivalent to an ensemble Kalman filter analysis which only utilizes observations taken up to and including the analysis time. The n_{th} iteration utilizes observations taken n observing times past the analysis time. Only the first iteration requires the integration of a forecast model.

Previous studies using idealized ensemble data assimilation systems (e.g. Hamill and Snyder 2000) have shown that the flow-dependent background-error covariances they provide are most beneficial when there are relatively few observations, i.e. when the observing network is sparse. When observations are very dense, the background-error covariances are not as flow-dependent, and the improvement over schemes with static background-error covariances, such as three-dimensional variational assimilation (3DVar), is not as great. In addition, the computational cost of recently proposed ensemble-data assimilation algorithms (Houtekamer and Mitchell 2001; Whitaker and Hamill 2001) is directly proportional to the number of observations being assimilated. Therefore, ensemble-based data assimilation should be both more computationally feasible and provide the greatest benefit over current operational schemes in situations when observations are sparse. Reanalysis before the radiosonde-era (pre-1948) is just such a situation.

In a companion study presented at this conference, the feasibility of reanalysis using only surface observations with an operational 3DVar system was demonstrated. Here we present details of the EnSRS formulation, and some results with a low-order model. Results with a more realistic general-circulation model will be presented at the conference which demonstrate how this method can translate information provided by surface observations into the middle and upper troposphere much more effectively than schemes with static background-error covariances.

*Corresponding author address: Jeffrey S. Whitaker, NOAA Climate Diagnostics Center, R/CDC1, 325 Broadway, Boulder, CO 80305; email: jsw@cdc.noaa.gov

2. AN ENSEMBLE SQUARE-ROOT SMOOTHER

Whitaker and Hamill (2001) introduced the ensemble square-root filter (EnSRF) as an alternative to the ensemble Kalman filter (EnKF) which does not require that noise be added to the observations (e.g. Houtekamer and Mitchell 1998; Burgers et al. 1998). The basic idea is that the ensemble mean and for deviations from the ensemble mean are updated separately in such a way the the ensemble mean analysis and analysis covariance are consistent with that predicted by Kalman filter theory.

Following the notation of Ide et al. (1997), let $\mathbf{x}^{\mathrm{b}}$ be an m-dimensional background model forecast; let $\mathbf{y}^{\mathrm{o}}$ be an p-dimensional set of observations; let $\mathbf{H}$ be the operator that converts the model state to the observation space; let $\mathbf{P}^{\mathrm{b}}$ be the $m \times m$-dimensional background error covariance matrix; and let $\mathbf{R}$ be the $p \times p$-dimensional observation-error covariance matrix. The minimum error-variance estimate of the analyzed state $\mathbf{x}^{\mathrm{a}}$ is then given by the traditional Kalman filter update equation (Lorenc 1986),

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{b}} + \mathbf{K}(\mathbf{y}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}), \tag{1}$$

where

$$\mathbf{K} = \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{2}$$

The analysis error covariance $\mathbf{P}^{\mathrm{a}}$ is reduced by the introduction of observations by an amount given by

$$\mathbf{P}^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^{\mathrm{b}}(\mathbf{I} - \mathbf{K}\mathbf{H})^{\mathrm{T}} + \mathbf{K}\mathbf{R}\mathbf{K}^{\mathrm{T}} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^{\mathrm{b}}. \tag{3}$$

In ensemble data assimilation, $\mathbf{P}^{\mathrm{b}}$ is approximated using the sample covariance from an ensemble of model forecasts. For the rest of the paper, the symbol $\mathbf{P}$ is used to denote the sample covariance from an ensemble, and $\mathbf{K}$ is understood to be computed using sample covariances. Expressing the variables as an ensemble mean (denoted by an over-bar) and a deviation from the mean (denoted by a prime), the update equations for the EnKF may be written as

$$\overline{\mathbf{x}}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{b}} + \mathbf{K}(\overline{\mathbf{y}}^{\mathrm{o}} - \mathbf{H}\overline{\mathbf{x}}^{\mathrm{b}}), \tag{4}$$

$$\mathbf{x}'^{\mathrm{a}} = \mathbf{x}'^{\mathrm{b}} + \widetilde{\mathbf{K}}(\mathbf{y}'^{\mathrm{o}} - \mathbf{H}\mathbf{x}'^{\mathrm{b}}), \tag{5}$$

where $\mathbf{P}^{\mathrm{b}} = \overline{\mathbf{x}'^{\mathrm{b}}\mathbf{x}'^{\mathrm{bT}}} \equiv \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_{\mathrm{i}}'^{\mathrm{b}}\mathbf{x}_{\mathrm{i}}'^{\mathrm{bT}}$, n is the ensemble size, $\mathbf{K}$ is the traditional Kalman gain given by (2) and $\widetilde{\mathbf{K}}$ is the gain used to update deviations from the ensemble mean. Note that wherever an over-bar is used in the context of a covariance estimate a factor of $n-1$ instead of n is implied in the denominator, so that the estimate is unbiased. In the EnKF, $\widetilde{\mathbf{K}} = \mathbf{K}$, and $\mathbf{y}'^{\mathrm{o}}$ are randomly drawn from the probability distribution of observation errors (Burgers et al. 1998). This choice of $\mathbf{y}'^{\mathrm{o}}$ ensures that for an infinitely large ensemble, (3) will be satisfied exactly (Burgers et al. 1998). However, as pointed out by Whitaker and Hamill (2001), for a finite ensemble (3) will not be satisfied exactly, and the noise added to the observations acts as an extra source of sampling error, degrading the performance of the filter. In the EnSRF, $\mathbf{y}'^{\mathrm{o}} = \mathbf{0}$ and $\widetilde{\mathbf{K}}$ is given by

$$\widetilde{\mathbf{K}} = \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\left[(\sqrt{\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}+\mathbf{R}})^{-1}\right]^{\mathrm{T}}(\sqrt{(\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}+\mathbf{R})}+\sqrt{\mathbf{R}})^{-1} \tag{6}$$

(Andrews 1968). This choice guarantees that (3) is satisfied exactly. If $\mathbf{R}$ is diagonal, observations may be assimilated serially, one at a time (Gelb et al. 1974), and the above expression simplifies to

$$\widetilde{\mathbf{K}} = \left(1+\sqrt{\frac{\mathbf{R}}{\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}+\mathbf{R}}}\right)^{-1}\mathbf{K}, \tag{7}$$

where $\mathbf{R}$ and $\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ are scalars, and $\mathbf{K}$ is a vector of the same dimension as the state vector of the model. This was first derived by Potter (1964). Although (6) requires the computation of two matrix square-roots, the serial processing version (7) requires only the computation of a scalar factor to weight the traditional Kalman gain, and therefore is no more computationally expensive than the EnKF.

Cohn et al. (1994) introduced a fixed-lag Kalman smoother as a means of providing retrospective analysis capability in data assimilation. The basic equations for the lag-0 implementation are the same as those of the Kalman filter (equations (1) - (3) above). For lag $l > 0$,

$$\overline{\mathbf{x}}^{\mathrm{a}}_{k|k+l} = \overline{\mathbf{x}}^{\mathrm{a}}_{k|k+l-1} + \mathbf{K}_{k|k+l}(\overline{\mathbf{y}}^{\mathrm{o}}_{k+l} - \mathbf{H}_{k+l}\overline{\mathbf{x}}^{\mathrm{b}}_{k+l|k+l-1}), \tag{8}$$

where

$$\mathbf{K}_{k|k+l} = (\mathbf{H}_{k+l}\mathbf{P}^{\mathrm{ba}}_{k+l,k|k+l-1})^{\mathrm{T}}[\mathbf{H}_{k+l}\mathbf{P}^{\mathrm{b}}_{k+l|k+l-1}\mathbf{H}^{\mathrm{T}}_{k+l} + \mathbf{R}_{k+l}]^{-1}. \tag{9}$$

The subscript notation $m|n$ refers to a quantity at observation time m, which incorporates knowledge of all observations up to and including time n. Thus, the standard Kalman filter update equation (1), expressed in this notation, would be

$$\overline{\mathbf{x}}^{\mathrm{a}}_{k|k} = \overline{\mathbf{x}}^{\mathrm{b}}_{k|k-1} + \mathbf{K}_{k|k}(\overline{\mathbf{y}}^{\mathrm{o}}_{k} - \mathbf{H}_{k}\overline{\mathbf{x}}^{\mathrm{b}}_{k|k-1}), \tag{10}$$

where

$$\mathbf{K}_{k|k} = (\mathbf{H}_{k}\mathbf{P}^{\mathrm{b}}_{k|k-1})^{\mathrm{T}}[\mathbf{H}_{k}\mathbf{P}^{\mathrm{b}}_{k|k-1}\mathbf{H}^{\mathrm{T}}_{k} + \mathbf{R}_{k}]^{-1}. \tag{11}$$

The EnSRS gain, $\mathbf{K}_{k|k+l}$, involves $\mathbf{P}^{\mathrm{ba}}_{k+l,k|k+l-1}$ which is the forecast-analysis error cross-covariance matrix between the the background field used in the Kalman filter update equation for time $k+l$, and the lag $l-1$ Kalman smoother analysis for time k. In the formulation of Cohn et al. (1994), this quantity is computed by propagating $\mathbf{P}^{\mathrm{aa}}_{k+1-1,k|k+l-1}$ directly using the dynamical model. In the Monte-Carlo formulation proposed here, the relevant quantity can be computed directly from the ensemble via

$$(\mathbf{H}_{k+l}\mathbf{P}^{\mathrm{ba}}_{k+l,k|k+l-1})^{\mathrm{T}} = \overline{\mathbf{x}'^{\mathrm{a}}_{k|k+l-1}(\mathbf{H}_{k+l}\mathbf{x}'^{\mathrm{b}}_{k+l|k+l-1})^{\mathrm{T}}}. \tag{12}$$

Therefore, the dynamical model need only be used to create the first-guess ensemble for the $l = 0$ filter analysis. Following the methodology used in the EnSRF, the lag l ensemble mean analysis is computed using (8), (9) and (12). The lag l analysis for deviations from the ensemble mean is computed using

$$\mathbf{x}'^{\mathrm{a}}_{k|k+l} = \mathbf{x}'^{\mathrm{a}}_{k|k+l-1} - \widetilde{\mathbf{K}}_{k|k+l}\mathbf{H}_{k+l}\mathbf{x}'^{\mathrm{b}}_{k+l|k+l-1}, \tag{13}$$

where $\widetilde{\mathbf{K}}_{k|k+l}$ is defined so that the ensemble analysis-error cross-covariance is exactly equal to the value predicted by the theory of Cohn et al. (1994),

$$\mathbf{P}^{\mathrm{a}}_{k|k+l} = \mathbf{P}^{\mathrm{a}}_{k|k+l-1} - \mathbf{K}_{k|k+l}\mathbf{H}_{k+l}\mathbf{P}^{\mathrm{ba}}_{k+l,k|k+l-1}. \tag{14}$$

If observations are processed one at a time, $\widetilde{\mathbf{K}}_{k|k+l}$ is a straightforward extension of the lag $l = 0$ result,

$$\widetilde{\mathbf{K}}_{k|k+l} = \left(1+\sqrt{\frac{\mathbf{R}_{k+l}}{\mathbf{H}_{k+l}\mathbf{P}^{\mathrm{b}}_{k+l|k+l-1}\mathbf{H}^{\mathrm{T}}_{k+l}+\mathbf{R}_{k+l}}}\right)^{-1}\mathbf{K}_{k|k+l}. \tag{15}$$

Performing a lag l EnSRS reanalysis processing observations serially at the analysis times $t_i = t_0 + i\Delta t, i = 1,2,...I$ involves the following steps;

1. Perform n parallel EnSRF analyses for each $t_i, i = 1,2,...I+l$, using (4), (5), (2) and (7) to update each of the n ensemble members. At each step a forecast model is then integrated forward n times using each analysis as an initial condition. The background-error covariances needed to compute the Kalman gain are computed using the sample covariance.

2. Perform n lag 1 EnSRS analyses for the observation times $t_i, i = 1,2,...I+l-1$ using (8) and (13) to update the ensemble mean and deviations, respectively. The sample covariance between the background forecasts at time t_{i+1} and the filter analyses at time t_i from step (1) are computed via (12).

3. Perform n lag 2 EnSRS analyses for the observation times $t_i, i = 1,2,...I+l-2$, using the sample covariance between the background forecasts performed in step (1) at time t_{i+2} and the lag 1 EnSRS analysis at time t_i produced at step (2) to compute the Kalman smoother gain in (9).

4. Repeat the procedure in step (3) for lags 3 to l. Each step uses the background forecasts at the observation locations from step (1) and the EnSRS analyses produced at the previous step.

This is essentially an iterative process with the n_{th} iteration using observations up to and including n observing times past the observation time.

3. RESULTS WITH A LOW-ORDER MODEL

At the conference, results of experiments with a T47, 15-level dry GCM will be presented, in which surface "pseudo-observations", sampled from an integration of the same model, are assimilated using the algorithm just described. The effect of the flow-dependant error covariances and smoother lag on the quality of the analysis, especially in the middle and upper troposphere, will be emphasized. Here we present some preliminary results with a much simpler model, the 40-dimensional model Lorenz and Emanuel (1998). This model is governed by the equation

$$\frac{dX_i}{dt} = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F, \qquad (16)$$

where $i = 1,...,m$ with cyclic boundary conditions. Here we use $m = 40$, $F = 8$ and a fourth-order Runge-Kutta time integration scheme with a time step of 0.05 units. For this parameter setting, the leading Lyapunov exponent implies an error-doubling time of about 8 time steps, and the fractal dimension of the attractor is about 27 (Lorenz and Emanuel 1998). For our assimilation experiments, each state variable is observed directly, and observations have uncorrelated errors with unit variance. Observations are processed serially (one after another) and are assimilated every time step for 10000 time steps (after a spin-up period of 1000 time steps).

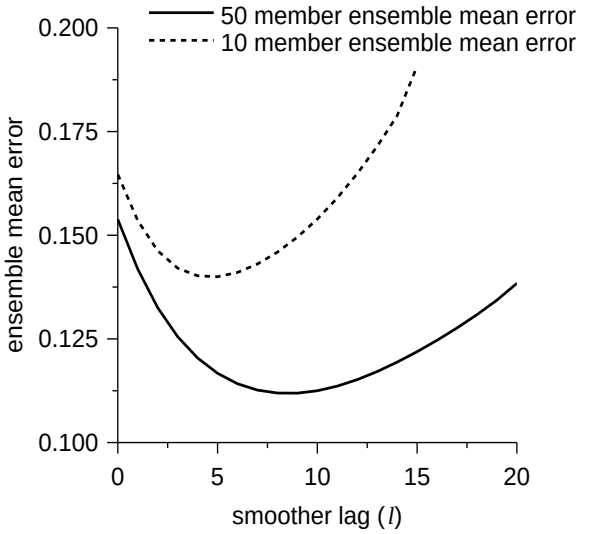

Figure 1: Ensemble mean error as a function of smoother lag for the 40-variable Lorenz model. Results are shown for a 10 (dotted) and a 50 (solid) member ensemble.

Figure 1 shows the ensemble mean error and ensemble spread as a function of smoother lag (l), for a 10 and 50 member ensemble. For a 10 (50) member ensemble, the EnSRS yields a 15% (27%) improvement in ensemble mean analysis error relative to the EnSRF for lag l = 5 (9). The differences between the errors in a 10 and a 50 member ensemble increase with filter lag. This indicates the sampling error in the estimation of the forecast-analysis error cross-covariance ($\mathbf{P}^{\mathrm{ba}}_{k+l,k|k+l-1}$) increases with l, so that a larger ensemble is needed to take advantage of observations farther removed from the analysis time. This is also consistent with the fact that, even with a 50 member ensemble, the quality of the analysis starts to degrade as the lag is increased beyond a certain point (about lag 9 for the 50 member ensemble and lag 5 for the 10 member ensemble). As discussed in Whitaker and Hamill (2001), sampling error can cause filter divergence in any ensemble data assimilation system, so some extra processing of the ensemble covariances is almost always necessary. The two techniques used here are distance-dependent covariance filtering (Houtekamer and Mitchell 2001; Hamill et al. 2001) and covariance inflation (Anderson and An-

derson 1999). For the results shown in Fig. 1, the parameters controlling the covariance filtering and inflation have been tuned to give the best filter (lag 0) analyses. These results indicate that since sampling error apparently increases with increasing lag, the optimal values of the covariance filter and inflation parameters are likely lag dependent.

References

Anderson, J. L. and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.

Andrews, A., 1968: A square root formulation of the Kalman covariance equations. *AIAA J.*, **6**, 1165–1168.

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

Cohn, S. E., N. S. Sivakumaran, and R. Todling, 1994: A fixed-lag Kalman smoother for retrospective data assimilation. *Mon. Wea. Rev.*, **122**, 2838–2867.

Gelb, A., J. F. Kasper, R. A. Nash, C. F. Price, and A. A. Sutherland, 1974: *Applied Optimal Estimation*. M. I. T. Press, 374 pp.

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

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**, in press.

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

— 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: operational, sequential, and variational. *J. Met. Soc. Japan*, **75 (1B)**, 181–189.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194.

Lorenz, E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399–414.

Potter, J., 1964: W matrix augmentation. M.I.T. Instrumentation Laboratory Memo SGA 5-64, Massachusetts Institute of Technology, Cambridge, MA.

Whitaker, J. S. and T. M. Hamill, 2001: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **129**, accepted, manuscript available at *http://www.cdc.noaa.gov/~jsw/pubs.html.*