7.4 ADAPTIVE ESTIMATION OF BACKGROUND AND OBSERVATION ERRORS WITHIN LOCAL ENSEMBLE TRANSFORM KALMAN FILTER

Eugenia Kalnay[1*], Hong Li[1] and Takemasa Miyoshi[2]
[1]University of Maryland, College Park, Maryland
[2]Numerical Prediction Division, Japan Meteorological Agency

1. INTRODUCTION

It is a common experience that OSSE experiments are more optimistic (give better forecast impacts) than real observation experiments. This is generally attributed to the fact that in OSSEs the model errors are neglected (or at least they are known). Another difference between OSSEs and real observation experiments, however, is that the observation error statistics are perfectly known in the OSSEs but not in real forecast experiments.

Recent diagnostic work within 3D-Var and 4D-Var (Desroziers et al, 2005, Talagrand 1999, Cardinali et al. 2004, Chapnik et al. 2006, and others) suggest that innovation and other statistics can be used to diagnose observation and background errors *a posteriori.* Miyoshi (2005) reported the use of the innovation statistics to estimate the background error inflation factor *online* within the Local Ensemble Transform Kalman Filter (Hunt et al, 2006, LETKF). Although the results were satisfactory this online estimation method relies on the assumption of the perfectly known observational errors.

Here we propose to adaptively estimate observational errors (for each type of instrument) and the inflation coefficient for the background error simultaneously within the LETKF. We will use the diagnostics in Deroziers et al. (2005) to estimate the background and observation errors, and adopt the method of Miyoshi (2005) to perform it adaptively within the LETKF, rather than *a posteriori*.

2. METHODS

2.1 Diagnosis of observation and background error statistics by Desroziers et al (2005)

Desroziers et al (2005) have explored the method to diagnose the observation and background error statistics in 4DVAR. They showed that for a system with perfectly tuned statistics:

$$< \mathbf{d}_{o-a}\mathbf{d}_{o-b}^T >= \mathbf{R} \quad (1)$$

$$< \mathbf{d}_{a-b}\mathbf{d}_{o-b}^T >= \mathbf{HBH}^T \quad (2)$$

R and B are the observation and background error covariance respectively. d_{o-a} (d_{o-b}) are the difference between the observation and analysis (forecast), and d_{a-b} is the difference between analysis and forecast.

Therefore they can estimate the observation and background error variances by:

$$\tilde{\sigma}_o^2 = \mathbf{d}_{o-a}^T\mathbf{d}_{o-b} / p = \sum_{j=1}^{p}(y_j^o - y_j^a)(y_j^o - y_j^b) / p \quad (3)$$

$$\tilde{\sigma}_b^2 = \mathbf{d}_{a-b}^T\mathbf{d}_{o-b} / p = \sum_{j=1}^{p}(y_j^a - y_j^b)(y_j^o - y_j^b) / p \quad (4)$$

p is the number of observations. The tildes in (3) and (4) indicate that this estimate is only approximate because it is based on the use of (generally incorrect) specified values for the observation and background error variances. In Desroziers et al 2005 equations (3) and (4) are estimations done offline, which allows the use of a large enough number of samples. They suggest schemes to estimate iteratively the true value of the observation and background variances from (3) and (4).

Unlike 3DVAR or 4DVAR, the background error covariance in LETKF is updated from the ensemble perturbations every analysis time step, rather than specified. However, the estimated background error covariance usually slightly underestimates the true forecast error covariance due to the limited number of ensemble members. Multiplicative (or additive) covariance inflation scheme is a simple and common method to avoid this problem (e.g., Anderson and Anderson 1999, Whitaker et al, 2006). However, *tuning* the inflation parameter is expensive, and there is no reason to think that it should be constant. We can adapt equation (4) to objectively estimate the inflation parameter using

* *Corresponding author address:* Eugenia Kalnay 3431 CSS, College Park, MD 20742; e-mail: ekalnay@atmos.umd.edu

$$< \mathbf{d}_{a-b}\mathbf{d}^T_{o-b} >= (1+\Delta)\mathbf{HP}^f\mathbf{H}^T \quad (5)$$

Here $\mathbf{P}^f$ is the time-dependent background error covariance and Δ is the inflation parameter required in LETKF. We can estimate Δ as

$$\tilde{\Delta} = \frac{\mathbf{d}^T_{a-b}\mathbf{d}_{o-b}}{sum(\mathbf{HP}^f\mathbf{H}^T)} - 1 \quad (6)$$

2.2 Adaptive estimation of covariance inflation

Miyoshi (2005) proposed another diagnostic based on observation-minus-background to adaptively estimate the inflation coefficient in EnKF. For a system with correctly specified inflation and observational error covariance,

$$< \mathbf{d}_{o-b}\mathbf{d}^T_{o-b} >= (1+\Delta)\mathbf{HP}^f\mathbf{H}^T + \mathbf{R} \quad (7)$$

Therefore, an estimate of the inflation can be obtained from

$$\tilde{\Delta} = \frac{\mathbf{d}^T_{o-b}\mathbf{d}_{o-b} - sum(\mathbf{R})}{sum(\mathbf{HP^fH})} - 1 \quad (8)$$

It is clear in order for this method to correctly estimate the inflation parameter it is necessary to correctly specify **R**. Miyoshi (2005) estimated the inflation parameter adaptively at each analysis time step. However, the number of samples available at each step is not enough, introducing a large sampling error. Therefore, he used a simple scalar KF approach usually used to postprocess model output (Kalnay, 2003, Appendix C) to accumulate past information and make the inflation gradually converge to the optimal value while still allowing for time variations. In addition, in order to avoid an unrealistically large sampling error that may occur and abruptly ruin the KF estimation, he imposed reasonable upper and lower limits in the estimation of Δ., e.g., $0.0 \le \Delta \le 0.2$.

To adaptively estimate Δ within EnKF, he calculated the inflation at each time step by using (8), and regarded it as an observed Δ^o for the current time step. Instead of directly using it as the final estimation for that time step, he used the simple scalar KF approach to best combine Δ^o and Δ^f, the value derived by persistence from the previous time step, to get a new estimate denoted as the analysis Δ^a:

$$\Delta^a = \frac{v^o\Delta^f + v^f\Delta^o}{v^o + v^f} \quad (9)$$

where v^f / v^o denotes the forecast/observational error variance weights for the scalar KF. The error variance of Δ^a is given by

$$v^a = (1 - \frac{v^f}{v^f + v^o})v^f \quad (10)$$

Assuming persistence as the forecast model for the inflation parameter, and allowing for some error in the "persistence forecast" (Kalnay, 2003, Appendix C), we have:

$$\Delta^f{}_{t+1} = \Delta^a{}_t \quad (11)$$

$$v^f{}_{t+1} = 1.03v^a{}_t \quad (12)$$

In this study, we use equation (3) to adaptively "observe" the observational error variance $\tilde{\sigma}^2_o$, and use either (6) or (8) to obtain the "observed" inflation $\tilde{\Delta}^o$ at each analysis time step. Then we use the simple scalar KF approach to make the best estimate of σ^2_o and Δ.

3. IMPLEMENTATION ON LORENZ 96 MODEL

First we test our approach with LETKF in Lorenz 96 40-variable model (Lorenz and Emanuel, 1998). The model time step is dt=0.025 and observations are assimilated with LETKF every 3dt. We integrate the model for 15000 steps to get the true state. At every analysis step we simulate the observations by adding the Gaussian random noise with standard deviation 1 to the 'true' state. Since the noise is uncorrelated, the true observation error matrix is diagonal, $\mathbf{R}_t = \sigma^2_{o(t)} = 1$.

We run the Lorenz 96 model for total 5000 analysis steps and only the last 2000 steps are used to calculate the analysis error and the online estimated σ^2_o and Δ

3.1 Perfectly specified observation variance

We first assume that the observation error variance is perfectly known, i.e., the specified value is $\sigma^2_{o(s)} = \sigma^2_{o(t)} = 1$. In this case we do not need estimate the R matrix, but we estimate inflation parameter adaptively using either equation (6) or (8) followed by the simple KF method. Table 1 shows that these two equations give similar results with estimated Δ around 0.1 and an analysis error of about 0.27. This experiment will serve as a benchmark for the later experiments where σ_o^2 is not perfectly specified anymore.

Table 1 Time mean of online estimated inflation parameter Δ and the corresponding analysis error, averaged over the last 2000 analysis steps in the case the true observational error variance is perfectly known.

Δ method	$\sigma_{o(s)}^2$	Δ	rms
Eq (6)	1	0.096	0.264
Eq (8)	1	0.101	0.270

3.2 Wrongly specified observation error variance

In reality we do not perfectly know the true value of the observation error variance, and the specified value used in the analysis is only an estimate. Our second experiment with the Lorenz96 model is to use an inaccurately specified $\sigma_{o(s)}^2$ which is ten times larger than the true $\sigma_{o(t)}^2$, and keep it constant throughout the analysis, as done in practice. With this wrongly specified σ_o^2, the LETKF will give too small weight to the observations. Therefore the observation information is not correctly used and the analysis is far from optimal. From table 2, we can see the analysis rms errors are much bigger than those from the benchmark run. In this experiment when we use equation (6) or (8) to estimate the inflation parameter, we have confined the estimated Δ to a reasonable range, $0.0 \le \Delta \le 0.2$. Without this restriction, the estimated Δ will be less than zero which makes the model blows-up.

Table 2: Time mean of online estimated inflation parameter Δ and the corresponding analysis error, averaged over the last 2000 analysis steps in the case the specified observation variance $\sigma_{o(s)}^2$ is ten times larger than the true $\sigma_{o(t)}^2$

Δ method	$\sigma_{o(s)}^2$	Δ	rms
Eq (6)		0.002	0.799
Eq (8)	10.0	0.008	1.088

3.3 Adaptive estimation of both the inflation and the observation error variance

We have seen that neither equation (6) nor (8) work when estimating the inflation parameter if the specified observation error information is wrong. In the third experiment, we estimate the observation variance and inflation simultaneously adaptively by using equation (3) and (6) or (8) followed by the simple KF method.

Table 3 shows that no matter how poorly the initial specified $\sigma_{o(ini)}^2$ is (ten times larger or ten times smaller than the true σ_o^2), equation (3) has the ability to correct it. The time mean of estimated σ_o^2 over the last 2000 analysis step is essentially the same as the true σ_o^2. Since R matrix is corrected, we can obtain a reasonable estimated inflation Δ which is about 0.1 for all the cases in table 3. Therefore the analysis rms errors are very small, and similar to those in the benchmark experiment.

Table 3, Time mean of online estimated observation variance σ_o^2, inflation parameter Δ and the analysis error, averaged over the last 2000 analysis steps in the case of initially wrong observation variance $\sigma_{o(ini)}^2$ but estimating and correcting it adaptively online.

R method	Δ method	$\sigma_{o(ini)}^2$	σ_o^2	Δ	rms
Eq(3)	Eq (6)	0.1	0.999	0.098	0.263
	Eq (8)		1.001	0.101	0.265
	Eq (6)	10.0	1.001	0.097	0.266
	Eq (8)		0.999	0.100	0.265

4. IMPLEMENTATION ON SPEEDY MODEL

In this section we will test our approach on a more realistic model. The SPEEDY model (Molteni 2003) is a recently developed atmospheric general circulation model (AGCM) with simplified physical parameterizations that are computationally efficient, but that maintain the basic characteristics of a state-of-the-art AGCM with complex physics.

The observations are obtained by adding zero mean normally distributed noise to the true state which is the two-month integration of the SPEEDY model from Jan 1 to Feb 28 in 1982. The observations are available on the model grid at every 4 grid points. The observed variables are zonal wind (u), meridional wind(v), temperature(T), specific humidity(q) and surface pressure (Ps) with error standard deviations of 1 m/s, 1m/s, 1K, 10^{-4} kg/kg and 100pa, respectively.

We double the true observational errors to get our first guess of the observational errors. Within LETKF, we estimate and correct these initially wrong observational errors every analysis time step (6-hour). We estimate the observational error covariance for each observed variable separately.

From the experiments of Lorenz96 model, we have seen as long as the observational error is corrected, we can get similar result whatever we use equation (6) or (8) to estimate the inflation parameter. Therefore here we will only test equation (8).

Fig.1 shows the online estimated observational errors for each observed variable. The experiment starts from wrong observational errors with 2 m/s for u and v, 2K for T, $2*10^{-4}$ kg/kg for q and 200 pa for Ps. After about 28 analysis time steps, i.e. one week, the estimated observational errors are already very close to their corresponding true values. As a result, we can get good analyses for all the variables. We compare the analysis errors from this estimating R experiment with those in which R is perfectly known, we find our approach for estimating R works very well. Although the RMS errors are a slightly larger than those from the perfect R case, they are already quite good (Fig. 2).

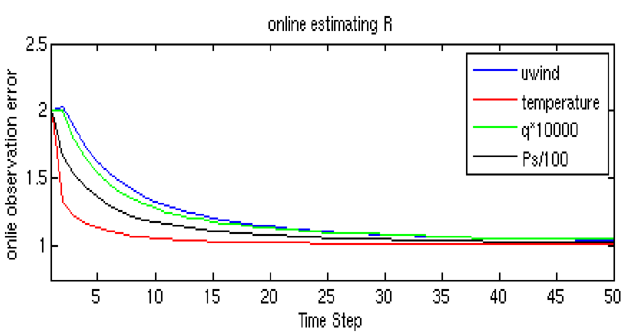

Fig.1 Time series of online estimated observational errors of u, T, q, and Ps for the first 50 analysis time steps (corresponding to 00z Jan 1 through 06z Jan 13, 2004)

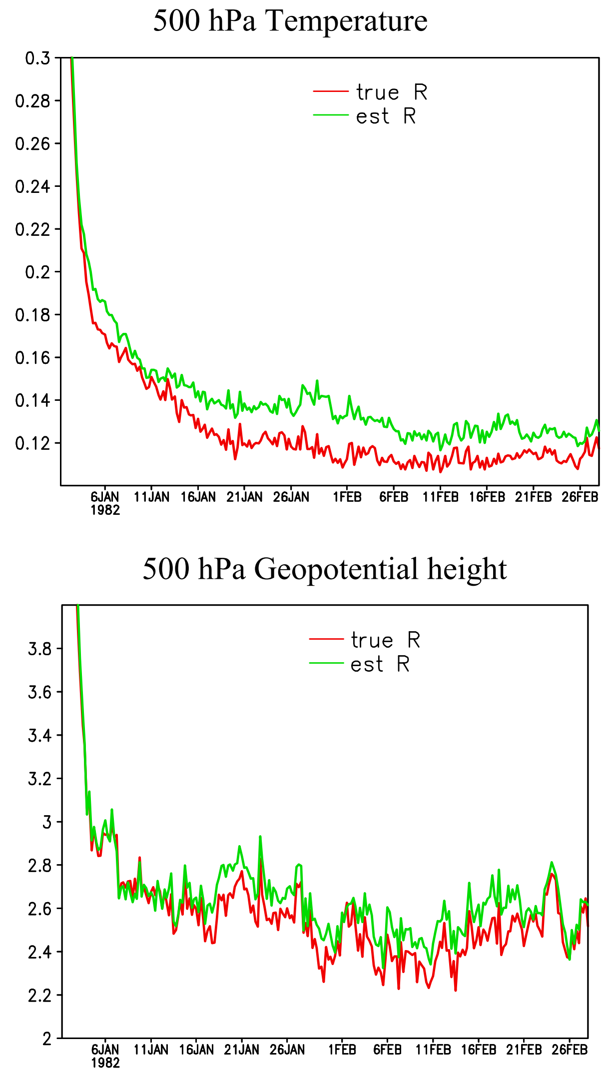

Fig.2 Time series of global averaged analysis RMS error of 500hPa temperature and geopotential height for January and February 2004 with the inflation estimated adaptively, in the cases of using a perfectly specified R (red) and using an initially wrong R but also estimating it adaptively (green).

5. CONCLUSIONS AND FUTURE WORK

The accuracy of an analysis system is dependent on appropriate statistics for observation and background errors. For ensemble based Kalman filter, tuning the inflation parameter is expensive. The online estimation method can objectively estimate the covariance inflation parameter but requires the appropriate information of observational errors. In this study, we estimate observational (for each type of instrument) errors and the inflation coefficient for the background error simultaneously within LETKF. The results show the online (adaptive) estimation of inflation parameter alone does not work without estimating the observational errors. Estimating both of them simultaneously our approach works well and can be applied to other ensemble based Kalman filters.

Currently, we are extending our approach to estimate off-diagonal terms in the observation error covariance, since so far we have only estimated the observational error variance. For some observations, like satellite retrievals, we have to account for the cross-correlations between observations. With our approach we hope to be able to adaptively estimate the correlations. We will test this first on the Lorenz96 and then the SPEEDY model.

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.

Cardinali, C, S. Pezzuli, and E. Andersson, 2004: Influence matrix diagnostic of a data assimilation system. *Q. J. R. Meteo. Soc*, **130**, 2767-2786.

Chapnik B., G. Desroziers, F. Rabier and O. Talagrand, 2006: Diagnosis and tuning observational error in a quasi-operational data assimilation setting. *Q. J. R. Meteo. Soc*, **132**, 543-565.

Desroziers G. and S. Ivanov, 2001: Diagnosis and adaptive tuning of osservation-error parameters in a variational assimilation. *Q. J. R. Meteor. Soc*, **127,** 1433-1452.

Desroziers G., L. Berre, B.Chapnik, and P.Poli, 2005: Diagnosis of observation, background and analysis error statistics in observation space. *Q. J. R. Meteor. Soc*, **131**, 3385-3396.

Hunt, B. R., 2006: Efficient Data Assimilation for Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter, In preparation.

Kalnay, E., 2003: Atmospheric modeling, data assimilation and predictability. Cambridge University Press, 341 pp.

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.

Miyoshi, T. 2005: Ensemble Kalman filter experiments with a primitive-equation global model, PhD Thesis at the University of Maryland.

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

Talagrand, O., 1999: a posteriori verification of analysis and assimilation algorithms. Proceeding of workshop on *Diagnosis of Data Assimilation Syst*em, November 1998, Reading, United Kingdom: ECMWF, 17-28

Whitaker, J. S., Hamil, T. H., Wei, X., Song, Y. C., and Toth, Z., 2006: Ensemble data assimilation with NCEP global forecast system. Submitted to *Mon. Wea. Rew*