6.1 APPLICATION OF LOCAL ENSEMBLE TRANSFORM KALMAN FILTER: PERFECT MODEL EXPERIMENTS WITH NASA FVGCM MODEL

Junjie Liu[1*], Elana Fertig[1], and Hong Li[1]
Istvan Szunyogh[1], Brian Hunt[1], E. Kalnay[1], Eric J. Kostelich[2], and Ricardo Todling[3]
[1]University of Maryland, College Park, MD, [2]Arizona State University,
[3]NASA-GSFC Global Modeling and Assimilation Office, Greenbelt, MD

1. Introduction

Ensemble Kalman Filter (EnKF) methods have been shown to be effective data assimilation schemes: Houtekamer et al. (2005) found the performance of an EnKF scheme to be comparable to that of an operational 3D-Var scheme when assimilating real observations into the CMC GEM grid point model. Using another EnKF scheme, Whitaker et al. (2004) obtained a better mid-troposphere reanalysis from surface pressure observations than with the NCEP 3D-Var. The Local Ensemble Kalman Filter (LEKF), another variant of the EnKF, was introduced in Ott et al. (2002; 2004) and was shown to be accurate and efficient when assimilating simulated observations of model variables on the NCEP GFS model in Szunyogh et al. (2005). In this study, the LEKF is replaced by a more efficient but equivalent implementation (Hunt, 2005), the Local Ensemble Transform Kalman Filter (LETKF). Here, we assimilate simulated model grid point observations and simulated rawinsonde observations into the NASA fvGCM model.

The LETKF algorithm is briefly described in Section 2, while its implementation on the fvGCM model is explained in Section 3. Section 4 present data assimilation results for the "perfect model" scenario. In these experiments, the NASA fvGCM is run for two months without assimilating observations to obtain a time series of "true" atmospheric state. Simulated noisy grid-point and rawinsonde observations of this truth are then assimilated with the LETKF. The performance of data assimilation system is evaluated by comparing the analysis state to the truth. Further verification of the scheme is obtained by comparing the performance of the LETKF to that of the NASA PSAS operational data assimilation scheme. A summary of the results which indicate the potential applicability of LETKF to assimilate real rawinsonde and AIRS observations is presented in Section 5.

[1*] *Corresponding author address, 2113 CSS, College Park, MD. Email: jjliu@atmos.umd.edu*

2. Local Ensemble Transform Kalman Filter

Data assimilation techniques find the best estimate of the state of the atmosphere by combining information from observations and previous forecasts, called the background. In ensemble data assimilation techniques, the background information is derived from an ensemble of k forecasts, the ith of which is denoted by $\mathbf{x}^{b(i)}$. The background state, $\bar{x}^b$, is estimated by the ensemble mean of these k forecasts,

$$\bar{x}^b = k^{-1}\sum_{i=1}^{k} x^{b(i)}$$

Similarly, the background error covariance matrix, P^b, is estimated by the sample covariance of the ensemble forecasts, that is,

$$P^b = (k-1)^{-1} X^b X^{bT}$$

where T denotes the matrix transpose and X^b is the matrix of ensemble perturbations with the ith column given by . $X^{b(i)} = x^{b(i)} - \bar{x}^b$

Following the derivation of Hunt (2005), with the above formulation of the background state, it can be shown that the Kalman Filter equations yield the following analysis equation:

$$\bar{x}^a = \bar{x}^b + X^b \tilde{P}^a Y^{bT} R^{-1}(y^o - \bar{y}^b) \quad (1)$$

$$P^a = X^b \tilde{P}^a X^{bT} \quad (2)$$

$$\tilde{P}^a = \left[(k-1)I + Y^{bT} R^{-1} Y^b\right]^{-1} \quad (3)$$

In these equations, $\bar{x}^a$ is the mean analysis state and P^a is the corresponding error covariance matrix and y^o is the vector of observations. $h(\cdot)$ is the operator mapping the model state variables into observation space and the ith background ensemble member in observation space is

$y^{b(i)} = h\left(x^{b(i)}\right)$ Thus, $\bar{y}^b = k^{-1}\sum_{i=1}^{k} y^{b(i)}$ is the mean background state in observation space. Finally, Y^b is the matrix of ensemble perturbations in observation space: that is, its *i*th column is given by $Y^{b(i)} = y^{b(i)} - \bar{y}^b$.

Similarly to the Ensemble Transform Kalman Filter of Bishop et al. (2001), the LETKF assumes that the analysis error covariance is in the space spanned by the background ensemble. Thus, the error covariance takes the form

$$P^a = (k-1)^{-1} X^a X^{aT} \quad (4)$$

where the *i*th column of the analysis ensemble perturbation matrix is given by the difference between the *i*th analysis ensemble member, $\boldsymbol{x}^{a(i)}$, and the analysis mean state, $\bar{x}^a$. From the above Kalman Filter equations, it is apparent that the matrix of ensemble perturbations is given by

$$X^a = X^b\left[(k-1)I + Y^{bT} R^{-1} Y^b\right]^{-1/2} \quad (5)$$

The above formulation for the perturbation matrix can be used to find the *i*th analysis ensemble member by $x^{a(i)} = X^{a(i)} + \bar{x}^a$.
LETKF is advantageous as it provides an optimal update for each ensemble member. Solving all the equations in the space spanned by the background ensemble members greatly reduces the computational cost involved in solving the Kalman Filter equations. Rather than solving for the entire high-dimensional state vector, the LETKF solves matrix equations of the same dimension as the relatively small number of ensemble members, *k*.

In the LETKF, following the concept of Ott et al. (2004), the state at each grid point is updated using the background and observations only within a distance from that grid point. This localization greatly reduces the cost of the assimilation by reducing the dimension of the LETKF matrix equations. Furthermore, because the state is updated independently at each grid point, the LETKF can process each grid point in parallel. Another benefit of the localization is that it filters spurious correlations between distant locations.

3. Implementation on fvGCM model

3.1 Forecast model

The NASA fvGCM is an operational weather forecasting model. The version employed in our experiments has 72 zonal grid points, 46 meridional grid points, and 55 vertical levels. We note that this resolution is much coarser than that used operationally, but it allows for a large number of numerical experiments on our available computational resources.

3.2 Observations

The observations are obtained by adding zero mean, Gaussian distributed noise with a prescribed variance to the true state.

Two sets of observations are generated to test the performance of LETKF with the fvGCM. The first set of observations is generated at model grid points. For this set of observations, the observed variables include the model variables of surface pressure, scaled potential temperature, zonal wind and meridional wind. The standard deviation of the observational errors for these model variables are 1hPa, 0.04K (roughly 1K in temperature), 1.1m/s, and 1.1m/s respectively. The other set of observations simulate rawinsonde observations. These observations include values of zonal wind, meriodinal wind, and geopotential height at real rawinsonde locations. The standard deviation of these observations is the same as those used operationally in PSAS, the operational 3D-Var assimilation scheme of NASA.

3.3 Numerical experiments

In the assimilation experiments carried out with grid point observations, the initial analysis cycle begins at 1800 UTC on 16 December 2002, and the true state of 0000 UTC at 15 Jan 2003 is chosen as the initial analysis mean state. In the assimilation experiments carried out with simulated rawinsonde observations, the initial analysis cycle starts from 1800 UTC on 01 Jan 2003, and the initial analysis mean state is the PSAS analysis at the same time. The initial analysis ensemble members are obtained by adding normally distributed noise to the mean analysis. The standard deviation of the analysis ensemble perturbation is the same as the standard deviation of the observational noise.

We also test the sensitivity of the analysis results to the observational coverage. In these experiments 100%, 30%, and 11% percent of the grid points are observed. In the two partial coverage cases, the locations observed are randomly chosen from all grid points. At the

chosen locations, all variables are observed at all model levels every six hours. We also test the sensitivity of the analyses to the observational frequency. In these experiments, the grid point observations are assimilated at only every twelve hours. The number of ensemble members is 40 in all experiments presented here.

In order to compensate for sampling errors and the effects of nonlinearities in the evolution of the estimation errors, a multiplicative variance inflation scheme is used (Anderson and Anderson 1999). In the grid point observation assimilation experiments, 8% inflation is used at each level. In the rawinsonde observation assimilation, the inflation factor varies with the vertical levels, such that the inflation factor is increased at the highest model levels. The local patch size used for localization depends on the observational coverage, the local patch is larger when the observations are sparser.

4. Results

4.1 Sensitivity to the observational spatial coverage

For the NCEP GFS model, Szunyogh et al. (2005) showed that the Local Ensemble Kalman filter could perform well even when observations were assimilated at only 2% of the model grid points. To test the implementation of the LETKF on the NASA fvGCM model, we test the performance of the analysis for different observation densities. Fig. 1 shows the time evolution of the 500hPa zonal wind and temperature RMS error averaged over the globe. It shows that the RMS error of the zonal wind is more sensitive to the observation density than the temperature, while the spin-up time is shorter for the zonal wind. But for the zonal wind, the difference in the RMS errors between the 30% and 11% coverage cases is only 0.2 m/s after the initial transient period. The difference between the RMS error of the temperature analysis at the 30% and 11% observational coverage is even smaller. At 11% observational coverage, the analysis RMS error drops below the observation error after a longer spin-up time. This conclusion is consistent with the results of Szunoygh et al. (2005).

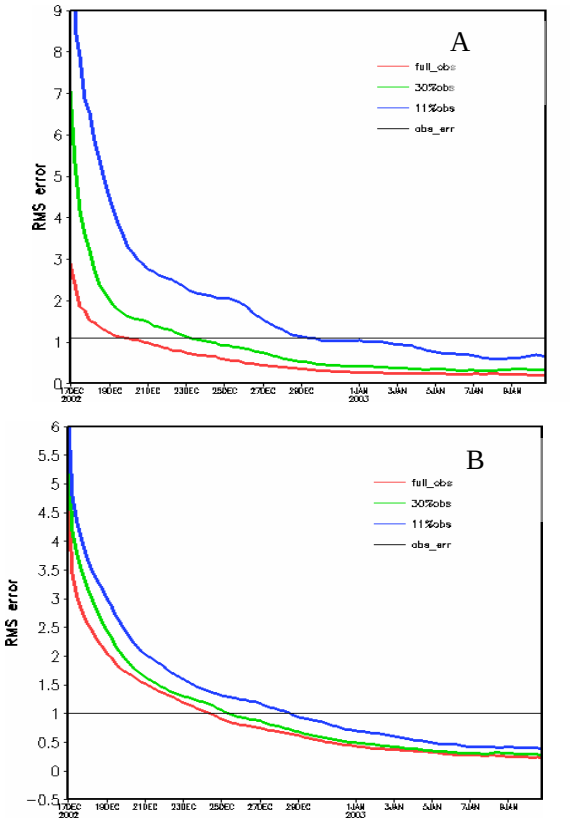

Fig. 1 Time evolution of the 500hPa analysis RMS error for zonal wind (A) and temperature (B) .The black line is the observation error, while the red, green and blue line show the analysis error for 100%, 30% and 11% observational coverage.

4.2 Sensitivity to analysis frequency

We test the sensitivity of the analysis results to the observational frequency at 30% observational coverage. Fig. 2 shows the time evolution of analysis RMS error for the 500hPa zonal wind (A) and temperature (B) for the 6-hour (red line) and 12-hour (green line) observational frequency. The zonal wind is more sensitive to the analysis interval than the temperature, but the performance of the LETKF with a 12-hour observation is competitive to that with a 6-hour observational frequency. The RMS error differs only by about 0.1 for zonal wind beyond the spin-up time. The RMS error for temperature is very similar for the two different analysis frequencies.

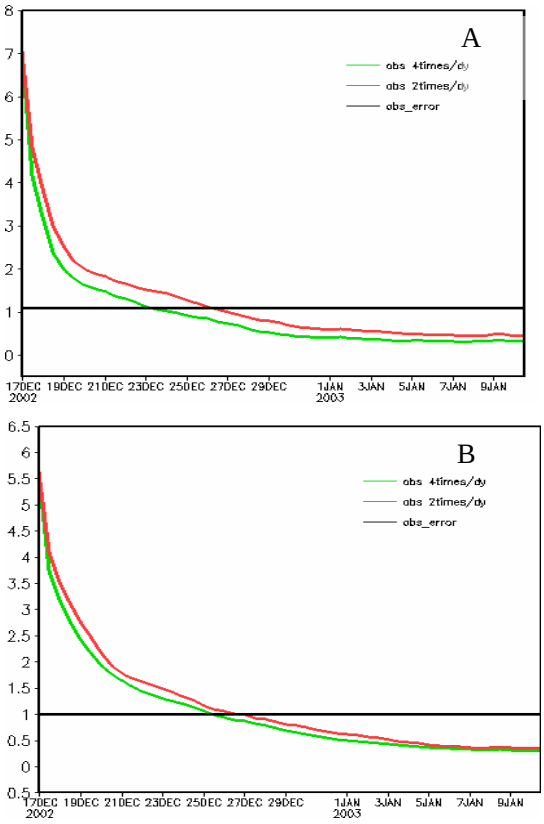

Fig.2 The global averaged 500hPa analysis RMS error (y-axis) as function of time (x-axis) for zonal wind (A) and temperature (B). The black line is the observation error, the red (green) line show the results for the 6-hour (12-hour) assimilation frequency,

The analysis errors strongly depend on the geographical location. The RMS errors in the zonal wind are the smallest over the mid-latitudes (25ºN/S-35ºN/S) at the lowest a few levels. The results for the 6-hour analysis frequency are almost uniformly better than that for the 12-hour analysis frequency, but the magnitude of the difference is small (Fig. 3B). The results for the temperature are similar (not shown).

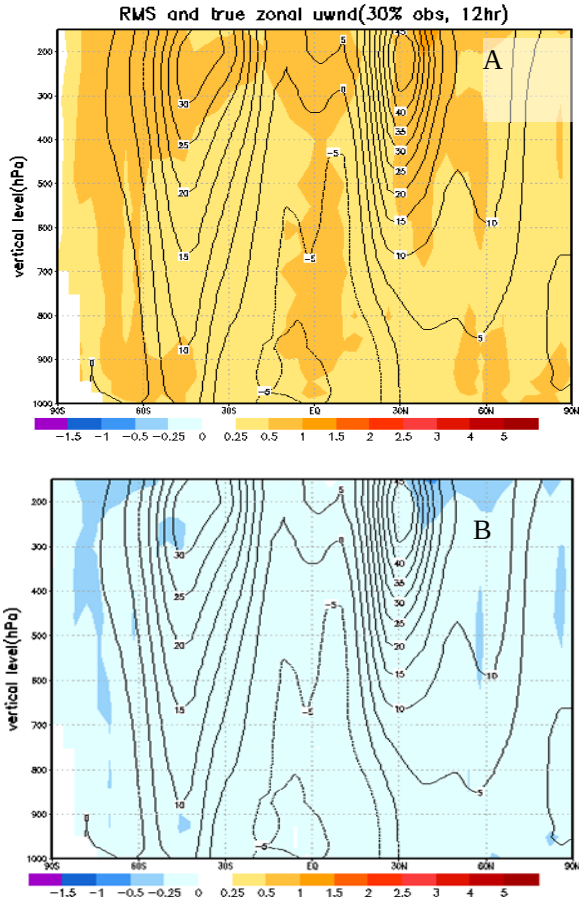

Fig. 3 Zonal average of the time mean zonal wind analysis error for 12-hour assimilation frequency (A), and the difference between the error at 6-hour analysis and 12-hour analysis frequency. The time average of the "true" zonal wind is also shown (contour).

4.3 Assimilation of simulated rawinsonde observations

The RMS error averaged over the Northern Hemisphere will be used to compare the performance of LETKF and PSAS. Compared with PSAS, the LETKF RMS errors are much smaller for both the wind and temperature beyond the spin-up time (Fig. 4). The RMS error difference between the two schemes is larger for the zonal wind than for the temperature (Fig.4). The RMS error difference is about 0.4 for zonal wind beyond the spin-up time, which is about one seventh of the observational error standard deviation (2.7m/s for 500hPa zonal wind observations). The spin-up time, about 6 days, is shorter for the zonal wind, which is consistent with the conclusion we drew from the experiments with the grid point

observations.

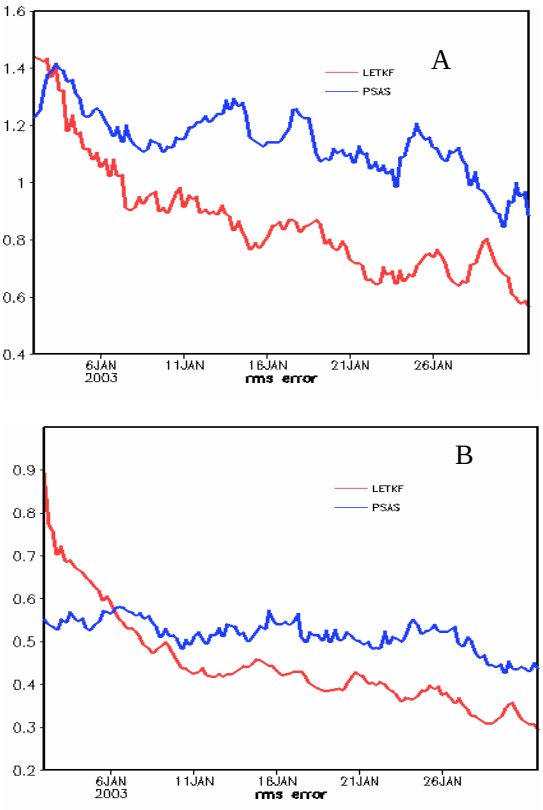

Fig.4 The RMS error for both LETKF (red line) and PSAS (blue line) of the 500hPa zonal wind (A) and 500hPa temperature (B).

Fig.5 shows the time mean (averaged over the last ten days analysis cycle (from 22 Jan 2003 to 31 Jan 2003) of the RMS error over the Northern Hemisphere. It is obvious that the LETKF analyses are clearly more accurate than the PSAS analyses for both the zonal wind and temperature at all levels between 1000hPa and 100hPa.

The RMS error of zonal wind does not change much up to 400hPa, and the RMS error in the temperature decreases with height up to 400hPa (Fig. 5). The changes of RMS error with height closely follow the changes in the observational coverage and observational error. Table 1 lists the observational error for different observation types and the number of observations at 0000 UTC at the different vertical levels. The zonal wind observation error increases with height up to 300hPa, while the observation coverage increases at the same time. Thus, the contributions of these two factors to the zonal wind RMS error cancel each other. The temperature is highly related with geopotential height observations. The observation error of geopotential height monotonically increases with height. Thus, the effect of the observation coverage dominates for the temperature RMS error variation.

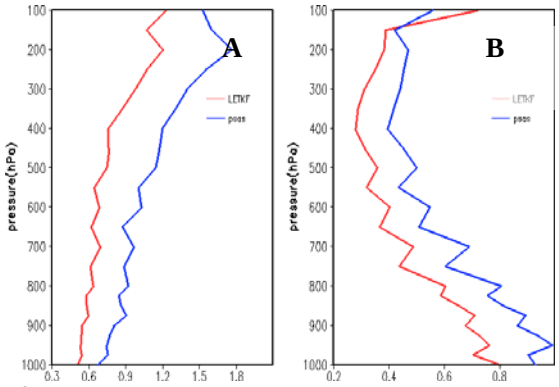

Fig. 5 Time mean of the RMS error of zonal wind (A) and temperature (B) averaged over Northern Hemisphere (the blue line is the error for PSAS, and the red line is the error for the LETKF) with simulated rawinsonde observation assimilation.

Unit (hPa)	U (m/s)	V (m/s)	H (m)	Total obs number
1000	2.0	2.0	5.4	607
850	2.2	2.2	5.6	1946
700	2.3	2.3	6.2	2057
500	2.7	2.7	8.6	2083
400	3.2	3.2	10.8	2038
300	3.4	3.4	12.8	1934
250	3.4	3.4	13.5	1529
200	3.3	3.3	14.5	1499
150	2.7	2.7	16.3	1460
100	2.7	2.7	19.3	1402

Table 1 Observation errors of different variables as function of vertical levels; Number of Observation at 0000 UTC as function of vertical levels (from NASA-PSAS).

Compared with PSAS, the LETKF analysis is more accurate almost everywhere except for the very high latitudes in the Southern Hemisphere (Fig. 6 B, Fig. 7B). Furthermore, the LETKF performs significantly better in the tropical region and mid-latitude of Southern Hemisphere than PSAS. The RMS error difference is larger for the

zonal wind (Fig. 6 B, Fig. 7B). The RMS errors are smaller in the mid-latitude of Northern Hemisphere and the tropical region for both variables (Fig. 6 A, Fig. 7A).

We recall that all results reported here were obtained by 40-member ensembles. We expect that the advantage of the LETKF over PSAS would be even larger if more ensemble members were added.

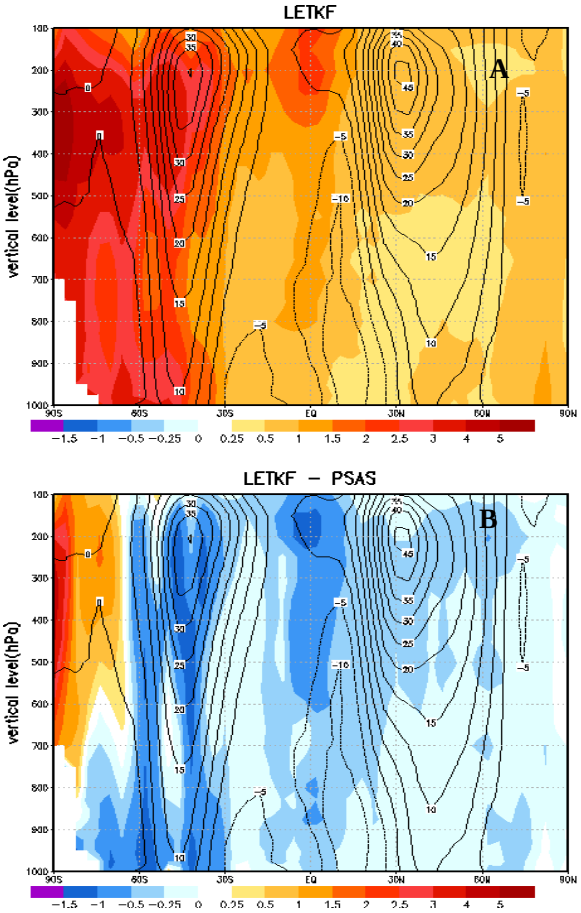

Fig. 6 Zonal average of the time-mean (average over the last 10 days) zonal wind analysis RMS error for the LETKF (A), and the RMS error difference between the PSAS and LETKF analysis is shown in B. The zonal average of the "true" time-mean zonal wind is shown by contours.

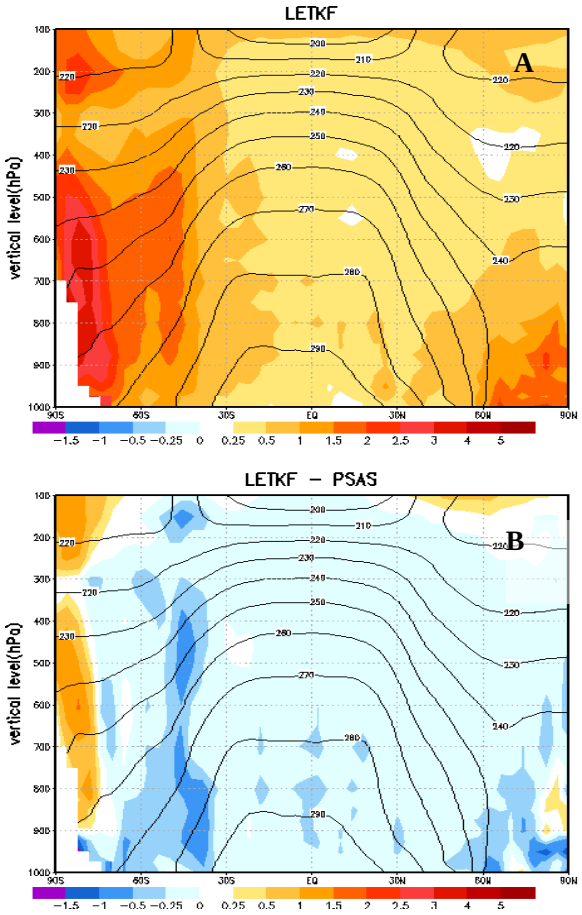

Fig. 7 Zonal average of the time-mean (average over the last 10 days analysis cycle) temperature analysis RMS error for both LETKF (A), and the RMS error difference between PSAS and LETKF is shown in B. The zonal average of the "true" time-mean zonal wind is shown by contour.

The temperature RMS error is relatively large in the Tropopause (Fig. 6A). The Tropopause is approximately at 200hPa in January. Geographical distribution of errors shows that the RMS errors are small over tropical and mid-latitude regions even at the Tropopause (Fig. 8A). The LETKF scheme outperforms PSAS most significantly in the tropical region (Fig. 8B). This result is especially encouraging since the tropical region is usually much more difficult to forecast and get good analysis.

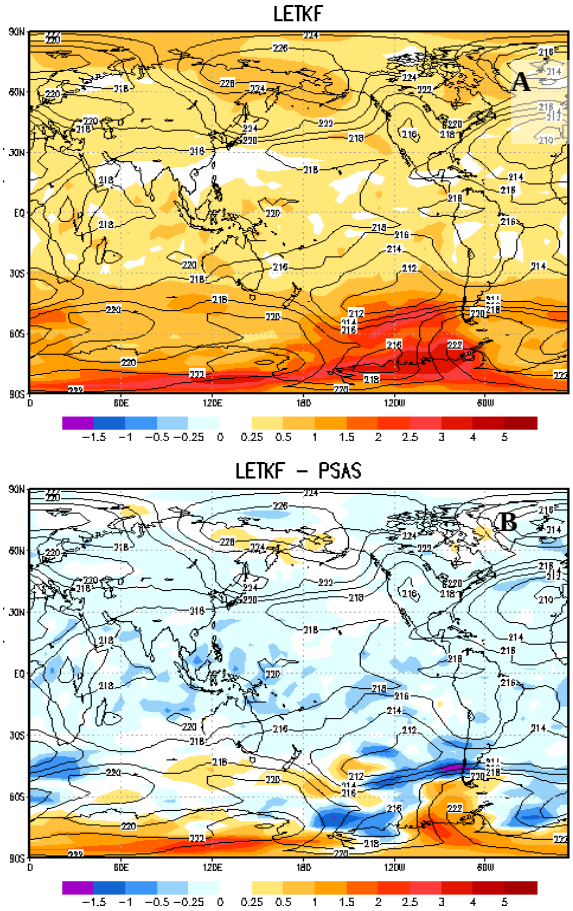

Fig. 8 Time-mean of the RMS error in the analysis of temperature component at the 200hPa (shades). The "true" time mean horizontal temperature at the same level is shown by contours.

5. Conclusion and discussion

The LETKF is an accurate and computationally efficient data assimilation scheme. It takes only 5 minutes on a 20 cluster of 2.8 MHZ PC processes to do an analysis at 30% observational coverage with 40-ensemble members.

Even in the case of sparse and less frequent observations, the LETKF can accurately assimilate grid point observations into the fvGCM model system. The sensitivity experiments show that the zonal wind analysis is more sensitive to the observational coverage and the observational frequency than the temperature. However, the spin-up time for temperature is longer than for the zonal wind.

With only 40 ensemble members, the LETKF scheme outperforms NASA PSAS analysis at almost geographical locations. The advantage of the LETKF over PSAS is the largest in the Tropics.

To further improve the performance of LETKF scheme, we are planning increase the number of the ensemble members to 60. With this ensemble member, we expect to obtain the better estimates of the background mean and background error covariances, which may compensate for the inadequate observation coverage in the Southern Hemisphere. We are also planning assimilate real rawinsonde observations and AIRS data in the near future.

References

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

Bishop, C. H., Etherton, B. J., and Majumdar, S. J., 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, **129**, 420-436.

Hunt, B. R., 2005: Efficient Data Assimilation for Spatiotemporal chaos: a Local Ensmeble Transform Kalman Filter, *In Preparation.*

Houtekamer, P.L, Mitchell, H.L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., and Hansen, B., 2005: Atmospheric data assimilation with the ensemble Kalman filter: Results with real observations. *Monthly Weather Review*, **133**, 604-620.

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.

Ott, E., Hunt, B.R., Szuynogh, I., Corazza, M., Kalnay, E., et al., 2002: Exploiting local low dimensionality of the atmospheric dynamics for efficient Kalman filtering. arXiv:archive/paper **0203058** http://arxiv.org/abs/physics/0203058.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E.J., Kostelich, M., Corazza, M., Sauer, T., Kalnay, E., Patil, D. J., and Yorke, J. A., 2004: A Local Ensemble Kalman Filter for Atmospheric Data Assimilation. *Tellus*, **56A**, 415-428.

Patil, D., Hunt, B. R. Kalnay, E., Yorke, J. A. and Ott, E., 2001: Local low dimensionality of atmospheric dynamics. *Phys. Rev. Lett.*. 86, 5878-5881.

Szunyogh, I., Kostelich, E. J., Gyarmati, G., Patil, D. J., Hunt, B. R., Kalnay, E., Ott, E., and Yorke, J. A., 2005: Assessing a Local Ensemble Kalman Filter: Perfect Model Experiments with the NCEP Global Model. *Tellus*, **57A**, 528-545.

Whitaker, J. S., Compo, G. P., Wei, X., and Hammil, T.H., 2004: Reanalysis without radiosondes using ensemble data assimilation. *Monthly Weather Review*, **132**, 1190-1200.