J1.6 AN ANALYSIS OF THE IMPACT OF OBSERVATIONAL DATA ON ETKF-BASED ENSEMBLE PERTURBATIONS

Mozheng Wei*, UCAR at NOAA/NWS/NCEP, Camp Springs, Maryland
Zoltan Toth, NOAA/NWS/NCEP, Camp Springs, Maryland
Richard Wobus, SAIC at NOAA/NWS/NCEP, Camp Springs, Maryland
Yuejian Zhu, NOAA/NWS/NCEP, Camp Springs, Maryland
Craig H. Bishop, UCAR at Naval Research Laboratory, Monterey, CA
Xuguang Wang, The Pennsylvania State University, University Park, PA

1. INTRODUCTION

Ensemble Transform Kalman Filter (ETKF) (Bishop et al. 2001) is used to generate initial perturbations for global ensemble forecasts by using NCEP GFS model which is the base model for the current NCEP global ensemble forecast system. The initial perturbations in the current NCEP bred vector based global ensemble forecast system are rescaled from the forecast perturbations by a rescaling factor that does not take the observation information into account (Toth and Kalnay 1997). An ideal rescaling factor should produce optimal analysis perturbations based on ensemble forecasts from the previous cycle and observations. One of our aims is to make ensemble perturbations more independent and flow-dependent.

ETKF, which is a variant of ensemble based Kalman square-root filters (Tippett et al. 2003), blends the forecast perturbations with instantaneous real-time observation data that are used by NCEP data assimilation system. It produces analysis perturbations that are based Kalman filter theory in ensemble representation. The impact of the observations can be seen from the comparison of results from the ETKF based ensemble forecast systems with different observational data.

2. BASIC FORMULATION

The ETKF formulation (Bishop et al. 2001) is based on the Kalman filter with the forecast and analysis covariance matrices being represented by k ensemble forecast and k analysis perturbations, it is one of the solutions from Kalman filter theory (Tippett et al. 2003). More details can be found in (Bishop et al. 2001, Bishop 2003, Wang and Bishop 2002, 2003).

*Corresponding author address: Mozheng Wei, NCEP Environmental Modeling Center, 5200 Auth Road, Rm. 207,Camp Springs, MD 20746
E-mail: mozheng.wei@noaa.gov

Let

$$\mathbf{Z}^f = \frac{1}{\sqrt{k-1}}[z_1^f, z_2^f, \ldots, z_k^f] \tag{1}$$

$$\mathbf{Z}^a = \frac{1}{\sqrt{k-1}}[z_1^a, z_2^a, \ldots, z_k^a] \tag{2}$$

where $z_i^f = x_i^f - x^f$ and $z_i^a = x_i^a - x^a$ $(i = 1, 2, \ldots, k)$ are k ensemble forecast and analysis perturbations. In our experiments, x^f is the mean of k ensemble forecasts (x_i^f), while x^a is the analysis from data assimilation. The forecast and analysis covariance matrices are formed respectively $\mathbf{P}^f = \mathbf{Z}^f \mathbf{Z}^{f\,T}$ and $\mathbf{P}^a = \mathbf{Z}^a \mathbf{Z}^{a\,T}$

For a given set of forecast perturbations $\mathbf{Z}^f$ at time t, the analysis perturbations $\mathbf{Z}^f$ can be solved from the Kalman filter equation

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{P}^f \mathbf{H}^T (\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{P}^\mathbf{f} \tag{3}$$

where $\mathbf{H}$ is the observational operator, $\mathbf{R}$ is the observational error covariance matrix. Let $\mathbf{A}^f = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{Z}^f$, A singular value decomposition of $\mathbf{A}^f$ is given by $\mathbf{A}^f = \mathbf{U}\Gamma^{1/2}\mathbf{C}^T$.

The ETKF solution is $\mathbf{Z}^a = \mathbf{Z}^f\mathbf{T}$, here $\mathbf{T} = \mathbf{C}(\Gamma + \mathbf{I})^{-1/2}$, $\mathbf{C}$ contains orthonormal right singular vectors and Γ is a diagonal matrix containing squared singular values of $\mathbf{A}^f$.

The analysis perturbations produced by ETKF in this way are not centered around the analysis, since $\sum_{i=1}^{k} z_i^a \neq 0.0$. A simple transformation that will preserve $\mathbf{P}^a$ and center the analysis perturbations around the analysis is the simplex transformation (Wang and Bishop 2003). $\mathbf{C}^T$ is one of the solutions of this transformation. Hence $\mathbf{Z}^a = \mathbf{Z}^f\mathbf{T}\mathbf{C}^T$ will be used as our initial analysis perturbations for the next cycle forecasts.

Since the number of ensemble members is too small comparing with the dimension of model state space, the analysis covariance is greatly underestimated. It is necessary to inflate the analysis perturbations. The inflation factors we used is based on the maximum likelihood on-line estimate (Dee

1995). Details can be found in Bishop 2003, Wang and Bishop (2003).

3. OBSERVATIONS

The experiments run from 12/31/2002 to 02/17/2003, our studies are focused on the 32 days period from 01/15/2003 to 2/17/2003. The observations used are from the conventional set in NCEP global data assimilation system. The conventional data set contains mostly rawinsonde and aircraft data. There are some wind data from satellites. The ensemble is cycled every 6 hours in accordance with the NCEP data assimilation system with new observations coming every 6 hours. The number of observations is changing from time to time. To have a snapshot of these observations, we show temperature and wind distributions at 00Z January 19, 2003 in Figs. 1 and 2. In Fig. 1 (a), only those temperature data below 500mb are shown horizontally. Fig. 1 (b) shows the numbers of temperature observations between different vertical levels.

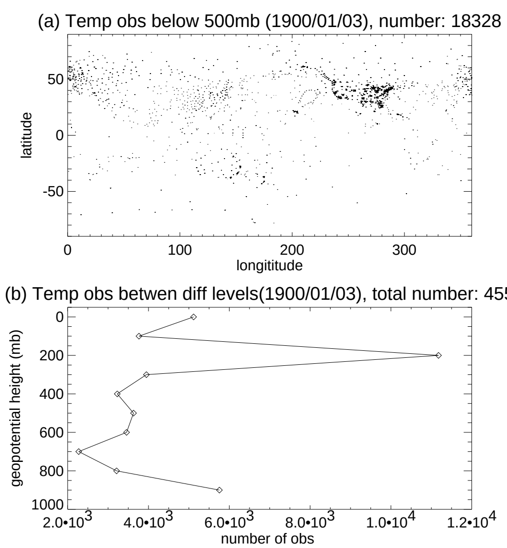

Figure 1: Temperature observation distributions at 00Z January 19, 2003. (a) the horizontal distributions of data below 500mb; (b) the vertical distribution.

For temperature, the data dense regions are in North America and Euro-Asia. However, the wind distribution is different. This is shown in Fig.2. The wind observations from satellites cover most of the tropic and both North and South Hemispheres.

4. VARIANCE DISTRIBUTIONS

An importance difference between the initial perturbations generated by ETKF based method and the current breeding method is that the former has taken the real-time observations into account. The roles played by the observations can be seen clearly from the observational space. We run 10 ensemble members, the same number as the NCEP operational ensemble system at the time of the experiments. The forecast and analysis covariance matrices in normalized observational space are $\mathbf{A}^f\mathbf{A}^{\mathbf{f}T}$ and $\mathbf{A}^a\mathbf{A}^{\mathbf{a}T}$ respectively, where $\mathbf{A}^a = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{Z}^a$. The variances in different eigen-directions are represented by the corresponding eigenvalues of the covariance matrices.

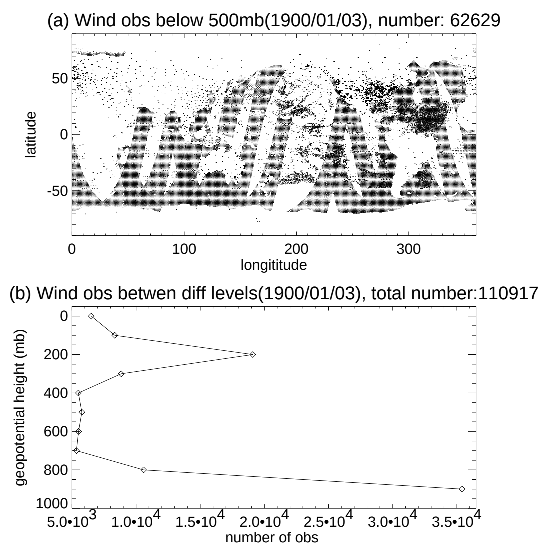

Figure 2: As in Fig.1, but for wind.

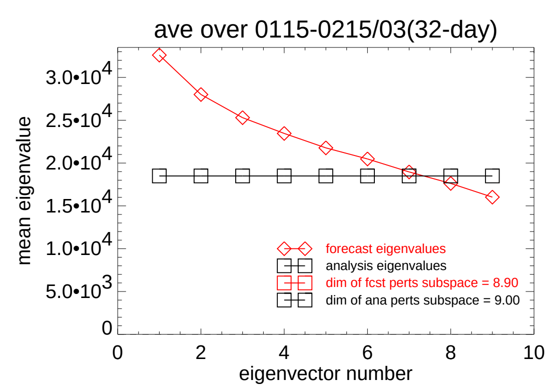

Figure 3: The averaged variance distributions along different eigen-directions of forecast and analysis covariance matrices in normalized observational space.

The red diamonds in Fig. 3 show the averaged eigenvalues of $\mathbf{A}^f\mathbf{A}^{\mathbf{f}^T}$ (6-hour forecast covariance matrix in normalized observational space) over the 32 days period from 00Z Jan. 15 to 00Z Feb. 15, 2003. There are only 9 independent directions out of 10 ensemble members. The dimension of subspace spanned by the 10 ensemble perturbations is 8.90 considering the variation of variances in different directions. In comparison with the bred vector based ensemble forecast system, the spectrum is more evenly distributed (not shown).

After ETKF analysis, the variances along different eigen-directions are almost equally distributed. This is shown in the black squares in Fig. 3. Note that the analysis variances are inflated to compensate the underestimations of small number ensemble representation. The dimension of the subspace spanned by the analysis perturbations is 9.00. The analysis perturbations are more independent than the forecast perturbations in the observational space.

5. IMPACT OF OBSERVATIONS

The main motivation behind using ETKF for ensemble forecasts is to bring the observations into play when we generate the initial perturbations for the ensemble forecasts for the next cycle, since the initial perturbations in ensemble forecast system should accurately represent the analysis errors in the right directions.

To test the impact of observations, we re-run the experiments with slightly different observation data at particular times. In the new experiments, we remove the Winter Storm Reconnaissance (WSR) data at 00Z on Jan. 19, 26, 31 and Feb. 01, 03, 08, and 09 2003. Details about 2003 WSR data can be found at http://wwwt.emc.ncep.noaa.gov/gmb/targobs/ target/wsr2003.html.

Each experiment starts from the exact initial condition as the original experiments at the previous cycle (i.e. 6 hours earlier). The new analysis perturbations on these 7 days at 00Z without WSR data will be compared those with the WSR data. On each day at 00Z, there are about 20 observations. Thus in each of the 7 cases, the difference between the experiments without and with WSR data will exactly reflect the impact of only 20 observations. The average results of these 7 cases are shown in Fig. 4.

Fig. 4 shows the differences between the two experiments without and with WSR data for rms analysis perturbations of temperature and wind (Figs. 4 (a) and (b)). The differences of two experiments for the ratios between analysis and forecast perturbations are shown in Figs. 4 (c) and (d) for temperature and wind. The black crosses indicate the positions of WSR data. It is clear that when WSR data are removed, analysis perturbations are larger in the areas where the WSR data are not available. It is indicated that larger analysis errors are generated in these areas without the WSR data. In another word, the WSR data reduce the analysis errors. The results demonstrate the more observations will reduce more analysis errors. Note that in some other areas outside the WSR data region, primarily near equator, there are some noises. Convections near tropics are more active than other regions, any difference including slightly different initial conditions which might come from the global model integration scheme will amplify quickly.

6. CONCLUSIONS

The experiments carried out in this study have demonstrated that the ETKF based ensemble forecast system is influenced by the observation data. In comparison with most ensemble forecast systems at major forecast operation centers, including bred vector and singular vector based systems which do not use observations at all, the ensemble Kalman filter based system can take advantages of available observations and reduce the analysis errors. This is a clear improvement over the current generation of ensemble forecast systems. Our results show that adding more observations in one area will generally reduce the analysis errors in that part of the area. Hence in order to enhance the ensemble forecast performance and reduce the errors in more regions, observations should cover areas as large as possible.

From the ETKF theory, the analysis perturbations are orthogonal in the normalized observational space and the variances are very much evenly distributed along different directions. One can imagine that larger areas with observations will increase the dimension of subspace spanned by the analysis perturbations. Much improvement can be generated by using a lot of more observations in comparison with the current generation of ensemble systems where the forecast perturbations show some strong correlations (Wei and Toth 2003).

One should note that the number of ensemble members is too small compared with the model state space. Projecting the huge variances from a large state space onto such a small subspace spanned by the ensemble is a simplification. However the number of ensemble members that can be implemented is limited by computing resources. For a given number of ensembles, it is important to inflate the analysis variances properly. At present, how to correctly inflate the analysis variances remains a challenging research issue.

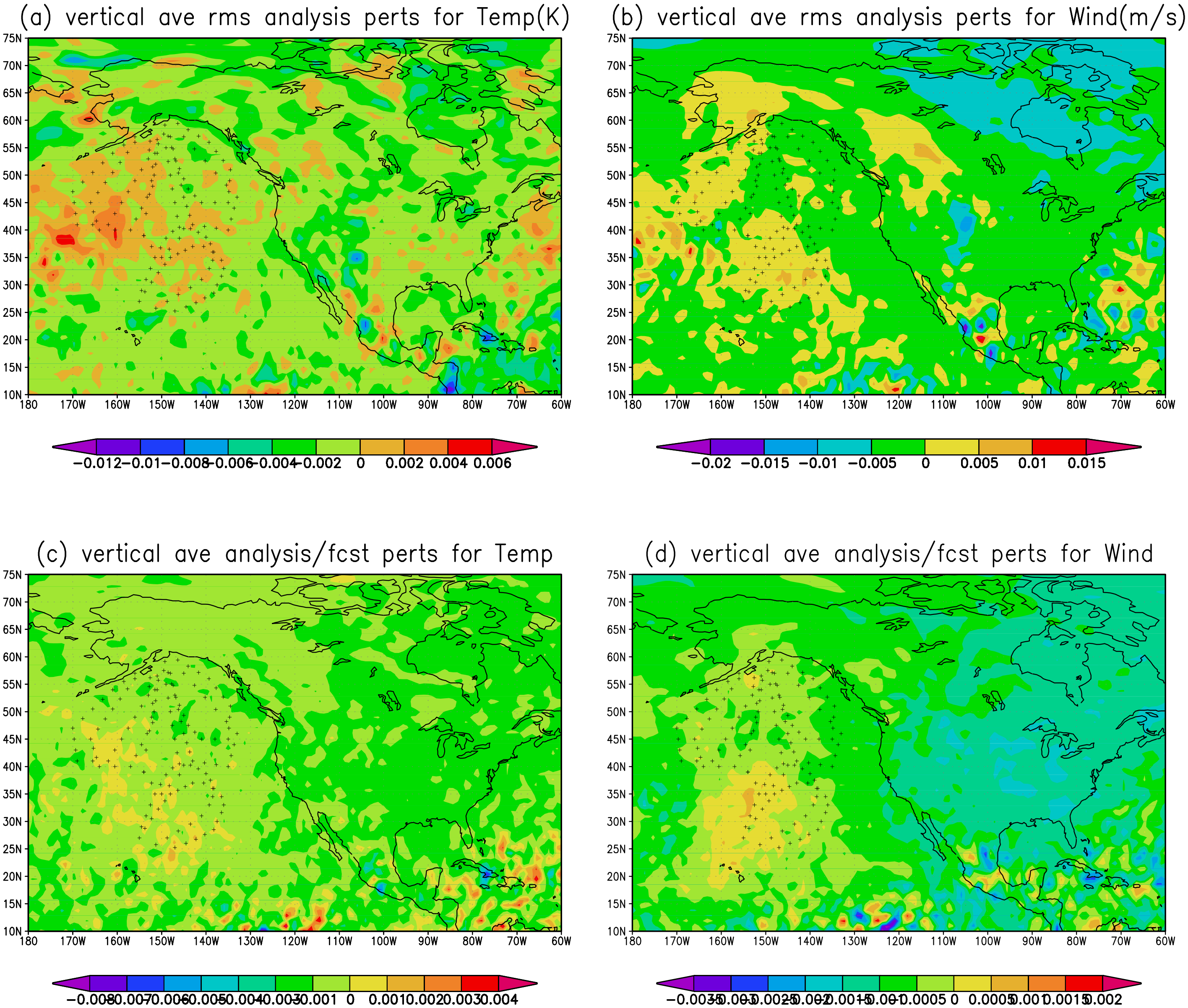

Figure 4: (a) and (b) are the differences of vertically averaged rms analysis perturbations for temperature and wind between the experiments without and with WSR data. (c) and (d) are the differences of experiments (without and with WSR data) of vertically averaged ratios between rms analysis and forecast perturbations.

Acknowledgements: We are grateful to many colleagues at NCEP/EMC for their help and useful discussion, particularly Lacey Holland, Russ Treadon, Wan-Shu Wu, Weiyu Yang, David Parrish, Jim Purser, Mark Iredell and Suranjana Saha.

7. 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**, 420-436.

Bishop, C.H. 2003: Amelioration of bias in the Ensemble Transform Kalman Filter. *Mon. Wea. Rev.*, submitted.

Dee, D. P. 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128-1196.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485-1490.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Wang, X., and C. H. Bishop, 2002: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci*, **60**, 1140-1158.

Wang, X., and C. H. Bishop, 2003: Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, submitted.

Wei, M., and Z. Toth, 2003: A new measure of ensemble performance: Perturbations versus Error Correlation Analysis (PECA). *Mon. Wea. Rev.*, **131**, 1549-1565.